Data Product Development Guide for Data Producers

Version 1.0, July 9, 2020

STATUS OF THIS MEMO

This memo provides information to the National Aeronautics and Space Administration (NASA) Earth Science Data Systems (ESDS) community. This memo describes a "Suggested Practice" and does not specify an ESDS standard of any kind. Distribution of this memo is unlimited.

CHANGE EXPLANATION

This is the original version of the document.

COPYRIGHT NOTICE

ABSTRACT

This Data Product Development Guide (DPDG) for Data Producers was produced for the Earth Observing System Data and Information System (EOSDIS) by the DPDG Working Group, one of the Earth Science Data System Working Groups (ESDSWGs) to aid in the development of NASA Earth Science data products.

The DPDG is intended for those who develop Earth Science data products and are collectively referred to as "data producers." This guide is primarily intended for developers of Earth Science data products derived from remote sensing data, and particularly for the development of Level 1B through Level 4 products. However, developers of related data products including Level 0 and 1A satellite data, airborne and ground-based data products will also find useful guidance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The National Aeronautics and Space Administration's (NASA's) Earth Observing System Data and Information System (EOSDIS) is a major capability in the Earth Science Data Systems (ESDS) Program [1]. EOSDIS Science Operations (data production, archive, distribution, and user services), which are managed by the Earth Science Data and Information System (ESDIS) Project [2], are performed within a system of interconnected Science Investigator-led Processing Systems (SIPSs) and discipline-specific data centers called Distributed Active Archive Centers (DAACs).

This Data Product Development Guide (DPDG) for Data Producers was produced for EOSDIS by the Earth Science Data System Working Groups (ESDSWGs) [3] under the supervision of the ESDIS Project to aid in the development of NASA Earth Science data products. We hope to update this document annually.

The DPDG is intended for those who develop Earth Science data products and are collectively referred to as "data producers" (see Appendix B). This guide is primarily intended for developers of Earth Science data products derived from remote sensing data, and particularly for the development of Level 1B through Level 4 products [4]. However, developers of related data products including Level 0 and 1A satellite data, airborne and ground-based data products will also find useful guidance.

There is an abundance of guides (e.g., standards, conventions, best practices, data format manuals) to direct developers in all aspects of designing and implementing data products. Moreover, some DAACs have developed guides for particular data producers and specific scientific communities [5] [6] [7] [8] [9] [10]. The DPDG aims to compile the most applicable parts of existing guides into an easy-to-follow document that logically outlines the typical development process for Earth Science data products. Emphasis has been given to standards and best practices formally endorsed by the ESDIS Standards Office (ESO) [11], findings from ESDSWGs, and recommendations from DAACs and experienced data producers.

Ultimately, the DPDG provides developers with guidelines for how to make data products that best serve end user communities—the primary beneficiaries of data product development.

The data products are assumed to be archived at a DAAC, and the data producers should work closely with the DAACs to which their products are assigned to obtain details not covered in this document.

The rest of the document is organized as follows. Section 2 covers the overall design and development process, including determination of requirements, design, implementation, testing, and independent review. Section 3 addresses selection of data formats. It covers recommended formats, other recognized formats, and data structures. The focus of Section 4 is on metadata. After a brief overview, recommended data product naming conventions, versioning, coordinate representation, and global as well as variable attributes are covered. Section 5 is a discussion of data chunking and compression. Section 6 provides a description of tools useful in data product development, including end-user tools with which the data producers should test data products to ensure their usability. Section 7 covers Digital Object Identifiers (DOIs) and actions data producers should take in collaboration with the DAACs. Section 8 describes the data publication process, once the data products are delivered to a DAAC. The organization of these sections is illustrated in the

flow chart in Figure 1 that shows the various steps in data product development and delivery. The numbers in parentheses in the figure indicate the sections where the individual steps are discussed. Following Section 8, a bibliography, a list of contributing authors and editors, abbreviations and acronyms, and a glossary are included. Finally, three appendices provide details of some tools, and important attributes that should be included in product metadata.

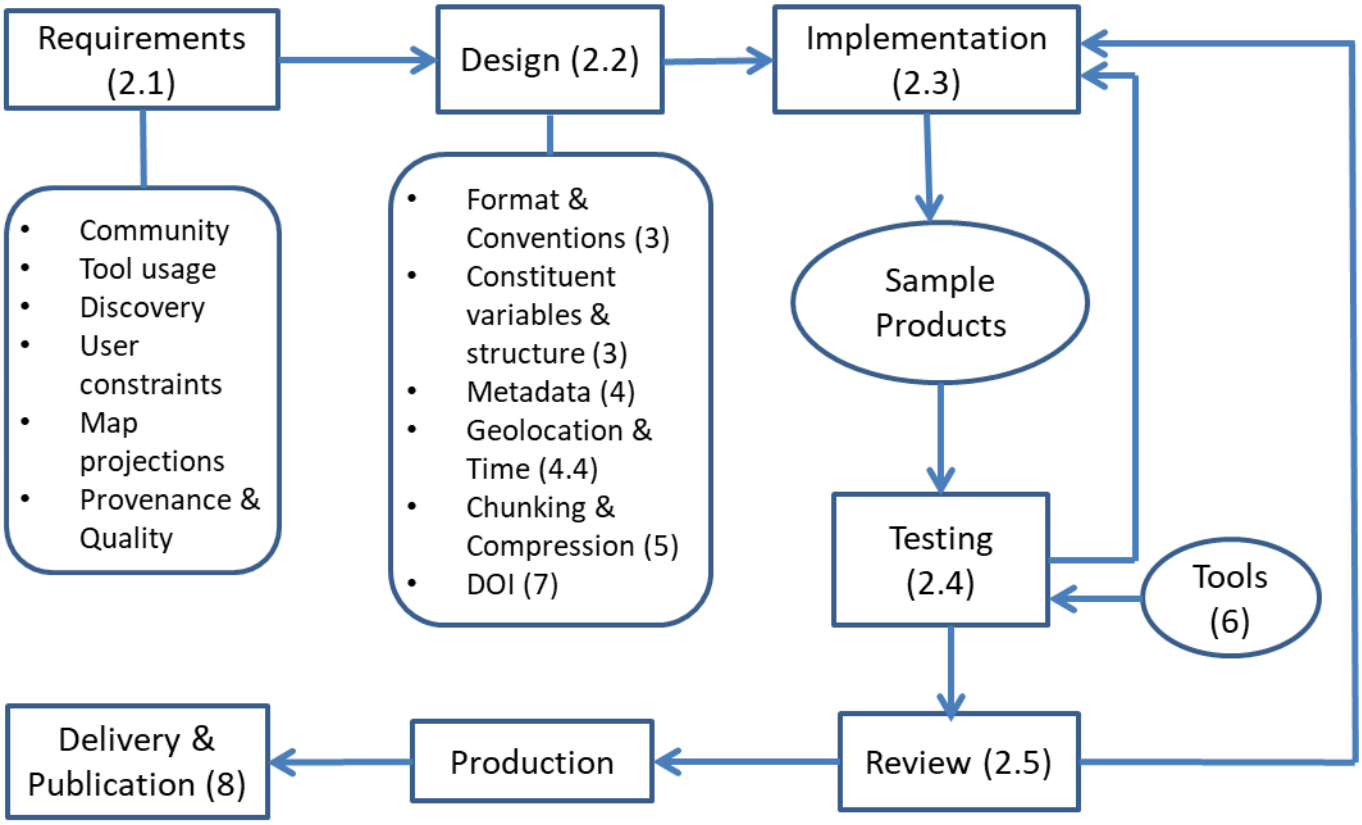

Figure 1. Flow of activities for data product development, production, and delivery (The numbers in parentheses in the figure indicate the sections where the individual steps are discussed).

2 DATA PRODUCT DESIGN PROCESS

For this guide, a data product (see Appendix B. GLOSSARY) is defined as a set of data files (see Appendix B) that can have multiple parameters, but compose a logically meaningful group of related data [12]. This concept is equivalent to a data collection (see Appendix B) in the Common Metadata Repository (CMR) [13], and is known colloquially as a dataset (see Appendix B).

Based on the Earth Observing System (EOS) heritage [14], standard Earth Science data products have the following characteristics:

- Peer-reviewed algorithmic bases.
- Wide research and application utility.
- Routinely produced over spatially and temporally extensive subsets of data.
- Available whenever and wherever the appropriate input data are available.

The following subsections discuss steps that can be applied in order to achieve good data product design.

2.1 Requirements: Determining User Community Needs

At the beginning of the design process, the key elements are to: 1) identify the expected user communities and 2) understand their needs regarding data formats, data structures, and metadata, in addition to what is required for data search and discovery. Developers can acquire this information by surveying the scientific literature, browsing predecessor data, holding and attending data applications workshops, and working with the DAAC that will archive the data product. While it is important to do this early in the product development process, it is equally important to remember that as Earth Science evolves to serve novel uses, user communities can change significantly. This is especially true in the case of long-term projects that involve several reprocessing episodes, where such changes could be accommodated in later versions of the products.

Once user communities are identified, the key questions to answer are:

- How can these data be used by the identified communities?
- Which tools and services will the community need to use the data?
- Are there common workflows applied to the data (e.g., subset >> quality filter >> re-grid)?
- What are the prevalent data formats, data and metadata standards and data structures used by the community?
- How does or should the community discover new data products (e.g., are there common keywords?)?
- What constraints does the community face (e.g., limited processor, disk storage, network bandwidth and timeliness)?
- What temporal and spatial resolutions, map projections and/or coordinate systems are commonly used in the community?
- What information on data provenance (data product history and lineage) and quality will the users need for their purposes?
- What other information do the users need to assess the suitability of the data?
- How should the data product be designed to make it useful to unexpected user communities (e.g., make the files self-describing)?
- What associated knowledge should be preserved [15] in addition to the data for the benefit of (long-term) future users, when data producers are no longer available for consultation?

2.2 Design: What Constitutes a Data Product Design

A data product design should address the following:

- Data format and associated conventions (Section 3).
- Identification and structure of constituent variables (Section 3.2).
- Metadata (Section 4).

- Representation of coordinates, especially geolocation and time (Section 4.4).
- Data chunking and internal compression (Section 5).

2.3 Implementation: Creating Sample Data Files

The data producer should create sample data files to support evaluation of the data product design. The more realistic the sample data, the more helpful it is for evaluating the usefulness and appropriateness of the design; however, even a data product populated with random values can be suitable for checking usability. Sample data coordinates should be accurately populated, as these are critical to the use of tools on a data product. Coding languages and various tools can be used to quickly create sample data files (Section 6).

2.4 Testing: Evaluating Sample Data Products

Testing of data products should be performed using the tools and services that user communities are expected to use (see Section 6). Typically, testing will identify structural issues of the data, such as missing or mis-identified variables and attributes. The data products should also be tested using compliance checkers (Section 6.2). Each data producer should consult with their DAAC regarding the available compliance checkers and how to use these tools. Ideally, an iterative approach should be followed—supply the data product to selected representatives of the expected user communities and DAACs, integrate feedback, re-test the product, and solicit additional feedback.

2.5 Review: Independent Evaluation of the Data Product

Soliciting external, independent evaluations will improve the usability of a data product. The following are recommendations for establishing and conducting such reviews throughout the product life cycle.

- Obtain reviews from the distributors of the data product (i.e., the relevant DAAC).
- To maintain objectivity, the evaluators should not be directly involved in the development of the data product. Ideally, evaluators should be representatives of the expected user communities.
- Perform multiple reviews during the development process to receive guidance well before the release of the product. Any resulting corrections should be documented.
- Feedback ought to be about both the content and quality of the data product. Four aspects of quality are defined in [16], namely scientific quality, product quality, stewardship quality and service quality. At this stage, the scientific and product quality are of primary concern. The content and quality should also help improve the applicability of the data product for specific uses.
- Responses to reviewers' comments should be documented and give the evaluators feedback on the disposition of their comments, clarifying any misunderstanding of the comments as needed.
- Reviews should address various aspects of the data product, including capabilities for search, access, exploration, analysis, interoperability, and usage.

- Reviews should verify that the data files and their variables have suitable names and enough supporting information to be understood.
- While data product user guides (i.e., those provided when users retrieve a data product or file) are essential to the usage of a data product, the data product ought to be self-describing (i.e., be embedded with information that describes the format and the meaning of the data).

It is also helpful to provide a way for the user community to provide feedback on the usability and quality of the data after the products are released, by including enough information in the data product metadata for the users to contact the data producer, or by working with the steward DAAC to glean relevant user feedback.

3 SELECTING A DATA PRODUCT FORMAT

Selecting a storage format for a data product involves weighing the advantages and disadvantages of applicable formats. Below are important considerations when selecting a format. Note that although the selected data format may not satisfy all users of a product, given sufficient user demand for a particular output format, EOSDIS can usually provide reformatting services to cater to a variety of preferences in the user community. The data producers are advised to consult their assigned DAACs to determine what reformatting services are available or can be implemented for satisfying the expected user community demand."

- Does the format provide for the widest possible use of the data product, including potentially new applications and research beyond the original intentions?
- Is the format widely used in the target user community for similar data or similar data analysis workflows?
- Is the format the same format as used for past long-term observations or models, and therefore can provide for or enable more efficient data processing and interoperability with those observations or models?
- Does the format enable efficient data analysis workflows on both global and local scales, and over long time periods as well as relatively near real time? Local (as opposed to global scale) applications will often require frequent subsetting, reprojection or reformatting of the data for combination with *in situ* point observations and physical models.
- Would it serve the user community better if the data were written in the same legacy format as was used for past or long-term observations for consistency, or in a format compatible with data from other agencies, e.g., the National Oceanic and Atmospheric Administration (NOAA) or the United States Geological Survey (USGS), to increase interoperability?
- Does the format support "self-describing" files, meaning that files contain metadata that describe the contents of the file?
- Does the format provide for efficient use of storage space (e.g., internal compression of data arrays), keeping file sizes practical for intended users, while minimizing the need to access multiple external files?
- Is the format supported by popular third-party applications (see Section 6) and programming environments, which could expand the user base and promote further development of tools and services?
- Is the version of the format supported by user community tools? The version of the format can be important: newly released formats may not be readable by the library and tool versions in use by the user community.

- Are resources available to support the format? This includes people who can develop and maintain libraries, tools and documentation for working with the format, and a commitment to long-term support of the format.
- Have optimizations, such as streaming, been considered (depending on the capabilities of the host facilities, and the envisioned use cases)?

Adherence to a well-defined set of specifications supports interdisciplinary data access, data sharing, research, and applications. The ESDIS Standards Office (ESO) [11] maintains approved standards, including those for data format, metadata, and data search and access [17]. For each of the items, the corresponding web page provides the strengths, weaknesses, applicability and limitations. Also, information is provided on deprecated standards and practices.

3.1 Recommended Formats

While a number of acceptable formats are listed by the ESDIS Standards Office (ESO) [18], the preferred format for EOSDIS data products is netCDF-4 (network Common Data Form Version 4) [19]. The netCDF-4 format uses the Hierarchical Data Format Version 5 (HDF5) [20] data storage format. Although files in netCDF-4 can in theory be written or read through HDF5 libraries and APIs (Application Programming Interface), inadvertent use of certain HDF5 features can render files unreadable by the rich ecosystem of netCDF tools, so we recommend use of the netCDF Application Programming Interface for all but the data format experts. The ESDIS Standards Office review of the netCDF-4/HDF5 File Format [19] lists the strengths, weaknesses, applicability and limitations of the format, which the reader may find useful.

Some of the advantages to using netCDF-4 include:

- Files are “self-describing”, meaning they allow for inclusion of metadata that describe the contents of the file. (These metadata should be inserted by the data producer to make the files truly self-describing). (See Appendix B).
- Supports many data storage structures, including multidimensional arrays and raster images.
- Includes access to useful HDF5 features, such as usability in HDF5 tools such as the HDF Product Designer (HPD) [21]
- Naturally accommodates hierarchical groupings of variables.
- Supports internal data compression.
- Supports several important programming languages and computing platforms used in Earth Science, so that the software and data platforms are language independent.
- Provides efficient input/output on high performance computing systems.

Also, a well-established standard called the Climate and Forecast (CF) Metadata Conventions (hereafter, CF Conventions – see Appendix B) [22] specifies a set of metadata that provide a definitive description of what the data in each variable represent and the spatial and temporal properties of the data. The CF Conventions were developed for netCDF; thus, they are sometimes referred to as “CF/NetCDF.” It should be noted that the CF Conventions were created for netCDF

files in which all data are located at root level (i.e., for files with no groups). Recently, the CF Conventions have been updated to include rules for files with group hierarchies [23].

3.1.1 NetCDF-4

A netCDF-4 file is organized into global attributes, dimensions, groups, group attributes, variables, and variable attributes. The global attributes provide general information regarding the file (e.g., author information, data product version, date-time range). Dimensions can represent: 1) spatio-temporal quantities (e.g., latitude, longitude, time); 2) other physical quantities (e.g., atmospheric pressure, wavelength); and 3) instrumental quantities (e.g., along track, cross track, waveband). A netCDF variable is an object that usually contains an array of numerical data. The structure of a variable is specified by its dimensions. The dimensions included at a given level in the hierarchy can be applied to variables at or below that level. Variable attributes provide specific information for each variable (e.g., coordinates, units, valid range). Groups can be created to contain variables with some commonality (e.g., ancillary data, geolocation data, and science data). Group attributes apply to everything in a group. Global attributes are attached to the root group.

Note that “dimensions” and “coordinates” are terms that should not be confused. For example, in a Level 2 (L2) swath file, the dimensions can be “along_track” and “cross_track,” while the corresponding coordinates can be “latitude,” “longitude,” and “time.” The coordinates for each variable are specified via the CF `coordinates`[1] attribute.

Data structures are containers for geolocation and science data. Guidance regarding swath structures in netCDF formats is provided in *Encoding of Swath Data in the CF Convention* [24]. The ESDSWG Dataset Interoperability Working Group (DIWG) has provided guidance regarding grid structures in netCDF-4 in [25] (Rec. 2.8-2.12) and [26] (Rec. 3.6). NOAA has provided a set of netCDF format templates for various types of data products [27], though these should be considered as informative, not definitive. Data producers can get guidance and samples from their DAAC. Earthdata Search [28] can also be used to acquire a variety of data in different formats and structures.

3.1.2 GeoTIFF

The GeoTIFF (Georeferenced Tagged Image File Format, *.tif) format is a georeferenced raster image that uses the public domain Tagged Image File Format (TIFF) [29], and is used extensively in the Geographic Information System (GIS) [30] and Open Geospatial Consortium (OGC) communities [31]. Although the types of metadata that can be added to GeoTIFF files are much more limited than with netCDF-4 and HDF5, the OGC GeoTIFF Standards Working Group is planning to work on reference system metadata in the near term. Both data producers and users find this file format easy to visualize and analyze, and so it has many uses in Earth Science. OGC GeoTIFF Standard, Version 1.1 is an EOSDIS recommended format [32]. Recently, a cloud-optimized profile for GeoTIFF has been developed to make retrieval of GeoTIFF data from Web Object Storage (see Appendix B, Glossary) more efficient [33].

[1] Words or phrases in this document that are colored `purple` indicate officially recognized attribute names.

3.2 Recognized Formats

In some cases, where the dominant user communities for a given data product have historically used other formats, it may be more appropriate to continue using those formats instead of the formats recommended above. If such formats are not already on ESO's list of approved data formats, they can be submitted to ESO for review and approval following the Request for Comments instructions [34].

3.2.1 ICARTT and Other ASCII Formats

The International Consortium for Atmospheric Research on Transport and Transformation (ICARTT) Version 2 format [35] arose from a consensus established across the atmospheric chemistry community for visualization, exchange, and storage of aircraft instrument observations. The format is text-based and composed of a metadata section (e.g., data source, uncertainties, contact information, and brief overview of measurement technique), and a data section. Although it was primarily designed for airborne data, the format is also used for other mobile and ground-based studies.

The simplicity of the ICARTT format allows files to be created and read with a single subprogram for multiple types of collection instruments and can assure interoperability between diverse user communities. However, ICARTT stores numbers using the American Standard Code for Information Interchange (ASCII), which can be far less efficient than binary formats such as netCDF-4, HDF5 or GeoTIFF regarding both data access and file size. Note that netCDF-4, HDF5 and GeoTIFF allow for the internal compression of variables, which results in a much smaller file size than any ASCII equivalent (see Section 5). Another disadvantage of ASCII is that print-read consistency is often lost. Two different programs that read the data could convert the ASCII to different binary numbers in the low order bits. This could complicate certain aspects of software engineering such as unit tests.

Other ASCII formats included in ESO's list of approved standards are: NASA Aerogeophysics ASCII File Format Convention, SeaBASS Data File Format, and YAML Encoding ASCII Format for GRACE/GRACE-FO Mission Data. In addition to ICARTT, these formats have been evaluated with respect to the ASCII File Format Guidelines for Earth Science Data [36] .

3.2.2 Vector Data and Shapefiles

The OGC GeoPackage is a platform-independent and standards-based data format for geographic information systems implemented as a SQLite database container (*.gpkg) [37]. It can store vector features, tile matrix sets of imagery and raster maps at various scales, and extensions in a single file.

OGC has standardized the KML (Keyhole Markup Language, *.kml) format that was created by Keyhole, Inc. and based on XML (eXtensible Markup Language) [38]. The format delivers browse-level data (e.g., images) and small amounts of vector data (e.g., sensor paths, region boundaries, point locations), but it is voluminous for storing large data arrays. KML supports only the geographic projection (i.e., evenly spaced longitude and latitude values), which can limit its usability. The format combines cartography with data geometry in a single file, which allows users flexibility to encode data and metadata in several different ways. However, this is a disadvantage to tool development and limits the ability of KML to serve as a long-term data archive format. OpenGIS KML is an

approved standard for use in EOSDIS. As noted in the recommendation, KML is primarily suited as a publishing format for the delivery of end-user visualization experiences. There are significant limitations to KML as a format for the delivery of data as an interchange format [39].

A Shapefile is a vector format for storing geometric location and attribute information of geographic features, and requires a minimum of three files to operate: the main file that stores the feature geometry (*.shp), the index file that stores the index of the feature geometry (*.shx), and the dBASE table that stores the attribute information of features (*.dbf) [40] [41]. Geographic features can be represented by points, lines, or polygons (areas). Geometries also support third and fourth dimensions as Z and M coordinates, for elevation and measure, respectively. Each of the component files is limited to 2 GB. Shapefiles have a number of limitations that impact storage of scientific data. “For example, they cannot store null values, they round up numbers, they have poor support for Unicode character strings, they do not allow field names longer than 10 characters, and they cannot store both a date and time in a field” [42]. Additional limitations are listed in the cited article.

3.2.3 HDF5

HDF5 is a modern data format designed to store and organize large amounts of data. NetCDF-4 (Section 3.1.1) and HDF-EOS5 (Section 3.2.4) are both built on HDF5. We strongly recommend that netCDF-4 be used for new Earth Science data products instead of HDF5, because Earth Science product files in HDF5 need to be compatible with netCDF-4 (i.e., readable via the netCDF-4 API) in order to work with netCDF-4 tools, and only very experienced product developers will be able to achieve this goal using the HDF5 API to write data.

3.2.4 HDF-EOS5

HDF-EOS5 is a specially developed data format for the Earth Observing System based on HDF5, which is widely used for NASA Earth Science data products and includes data structures specifically designed for Earth Science data.

HDF-EOS5 employs the HDF-EOS data model [43] [44], which remains valuable for developing Earth Science data products. The Science Data Production (SDP) Toolkit [43] and HDF-EOS5 library provide the API for creating HDF-EOS5 files that are compliant with the EOS data model.

In choosing between HDF-EOS5 and netCDF-4 (based on HDF5) with CF conventions, netCDF-4/CF is recommended over HDF-EOS5 due to the much larger set of tools supporting the format.

3.2.5 Legacy Formats

Legacy formats (e.g., netCDF-3, HDF4, HDF-EOS2, and plain ASCII) are those used in early EOS missions, though some missions continue to produce data products in these formats. Development of new data products or versions from early missions may continue using the legacy format, but product developers are encouraged to transition data to the netCDF-4 format for improved interoperability with data from recent missions. Legacy formats are recommended for use only in cases where the user community provides strong evidence that research will be hampered if the data formats are changed.

3.2.6 Other Formats

Some data products are provided by data producers in formats that are not endorsed by ESO, especially data collected during field campaigns. Producers of these data are not necessarily NASA-funded; thus, they are not under an obligation to conform to NASA's format requirements or could lack adequate resources to do so.

There are other formats that are currently evolving in the community, especially with cloud computing, big data, and analysis-ready data. As these are adopted and matured, ESO could consider them and recommend their use for NASA Earth Science data products. Such recommendations will be considered in future versions of this document.

4 METADATA

4.1 Overview

Metadata are information about data. As with the other aspects of data product development, it is helpful to consider the purpose of metadata in the context of how users will interact with the data and how metadata are associated with (i.e., structurally linked to) the data.

Metadata are essential for data management: they describe where and how data are produced, stored, and retrieved. Metadata are also essential for data search/discovery and interpretation, including facilitating the users' understating of data quality. A data producer has a responsibility to provide adequate metadata describing the data product on both the data-product-level and the file-level. The DAAC that will archive the product is responsible for maintaining the data-product-level metadata, but metadata can be augmented with details from the data producer. A subset of data-product-level and file-level (sometimes called granule-level) metadata that are relevant to data search/discovery is submitted to CMR by the DAAC(s).

Data-product- and file-level metadata are stipulated by the policies of the DAAC that will host the data. Although CMR has somewhat minimal requirements for metadata, requirements on EOSDIS data are expected to exceed those in order to render the data searchable and usable. The data producers should also follow best practices for including metadata within or associated with the individual data files that are submitted to the DAAC. Data producers should work with the DAAC(s) to determine the best approach for their products.

File metadata can be created for and associated with a data product through several methods. Software libraries, such as netCDF and HDF, make populating file metadata straightforward. Metadata can be assigned to any object within the file, or to the file as a whole. The two main categories of file metadata are "global attributes" and "variable attributes." Metadata can also be assigned to groups of data objects within a file. Global attributes are meant to apply to all information in the file, and can vary from file to file within a data product. However, to maximize the self-describing nature of a file a data producer can also include data product metadata, i.e. information that is repeated for all files, within a file. Data product metadata describe and are valid for data in each of the files constituting the data product. File-level metadata should be embedded in the file itself if using self-describing formats like netCDF. The DAAC(s) may require that the metadata be provided both embedded in files and as a separate metadata file. File-naming

conventions should make sure that the physically separate metadata are properly associated with the file to which they refer. Data product files may not contain all the available data product metadata (e.g., may not contain everything in related texts such as the Algorithm Theoretical Basis Document, ATBD), but they must contain enough metadata to enable data search and discovery and scientific analysis using tools capable of recognizing metadata standardized for interoperability based on recommendations and standards in this document and provided by the DAAC(s).

4.1.1 Data Product Search and Discovery

Most users will encounter metadata for the first time during the search and discovery process—when they are searching for data products that meet their needs. As mentioned above, the metadata that support this process are typically ingested into the CMR by the DAAC(s). These metadata are used either by the CMR search engine or by other search engines that harvest metadata (e.g., data.gov or Google).

Key success criteria for metadata during the discovery process include:

- Intelligible, descriptive data product names (Section 4.2).
- Precise temporal and spatial coverage (Section 4.4 and 4.5).
- Accurate and complete list of applicable Global Change Master Directory (GCMD) Science Keywords [45] as well as other keywords.
- Concise but readable description.

Note that because a given data product will be compared to thousands of other data products in the archives (approximately 32,000 in the EOSDIS catalog, the CMR), it is crucial to use standard names [45], especially for the platform (e.g., Nimbus-7), instrument (e.g., TOMS), and science (e.g., OZONE) keywords. Reference [46] provides links through which various categories of keywords can be downloaded in a variety of formats such as comma-separated variables - CSV). When the standard names (keywords) in the presently available list of GCMD are not directly applicable to a data product, the data producer is advised to follow the proper GCMD list update procedure [47]. Data producers should work closely with their assigned DAACs in selecting keywords for their products.

4.1.2 File Search and Retrieval

Once a user has chosen a data product to pursue, the user typically needs only some files, not the entire data product. Because metadata are standardized, data search engines, such as Earthdata Search for CMR [28], support the specification of spatial and temporal criteria (i.e., search filters). Therefore, it is best to precisely specify the spatial extent of a given file to limit “false positives” in search results. For example, a four-point polygon provides a more precise specification of spatial extent than a bounding box (Figure 2). Data producers should consult their assigned DAACs regarding methods the DAACs are currently using for specifying bounding regions before deciding whether a different approach is needed for their products.

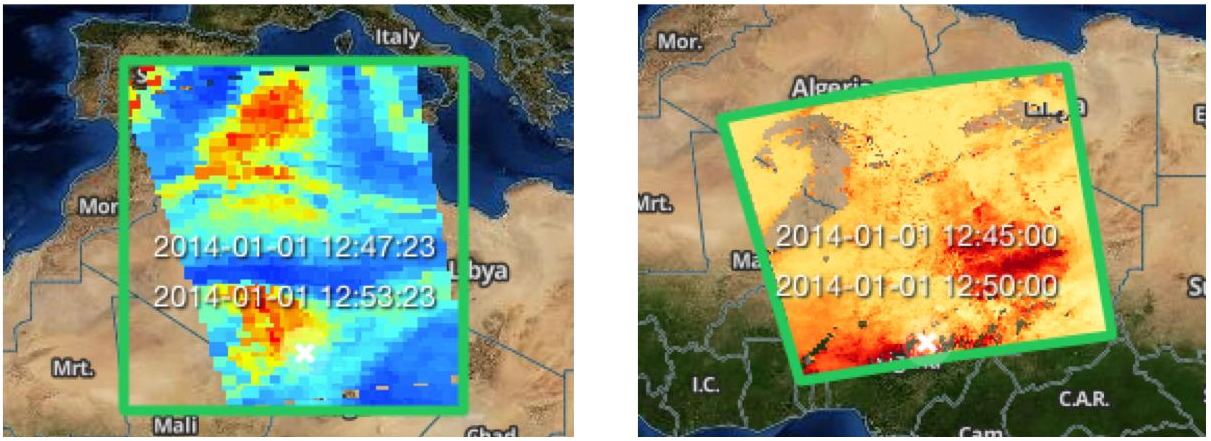

Figure 2. Screenshots from Earthdata Search for scenes from two Level 2 satellite data products.

Left: The green box represents the file outline, which is a bounding box provided by the spatial extent specified in the metadata. The bounding box results in large areas of "no data" in the southwest and northeast corners, where a user selection generates a false positive result. **Right**: True data coverage illustrated with a superimposed browse image and uses a four-point polygon with negligible areas where a user selection produces a false positive.

4.1.3 Data Usage

Finally, when the data are to be used, high-quality metadata are essential for data readability, not only for human users, but also for the software and other non-human users, such as decision support systems or models. Because files often contain metadata *per se*, the aim should be for data to be usable without modification by a user—a benchmark known as "Analysis-Ready Data" [48], the preparation of which typically requires:

- Coordinate information: spatial and temporal coordinates in standard form.
- Interpretation aids: units [26] (Rec. 3.2 and 3.3), fill (missing) value identification (see the DIWG Recommendations Part 3 [in preparation]), offset and scale factors [25] (Rec. 2.2, 2.5 and 2.6).
- Readable variable names, using standardized GCMD names and CF standard names, where applicable. For non-standard names, holding to consistent abbreviations and case conventions is advisable.

In addition, data producers should be mindful of other types of potentially useful documentation to facilitate the understanding of data, including but not limited to: provenance information, particularly identifiers of the data inputs and algorithm version, and pointers to documentation such as ATBDs, and data quality assessment(s). Such documentation should be produced as early as possible in the dataset production lifecycle.

Since the primary purpose of CMR is to support search and discovery, at present, not all the data usage information is ingested into CMR. However, as automated data system capabilities are evolving to provide higher levels of service for new data products, it is expected that more of the data usage information will need to be included in CMR.

4.2 Naming Data Products

The name of a data product is critical to its discovery in Earthdata Search and tools developed by DAACs. DAACs have rules for naming data products, so selection of long and short names should be a joint effort between producer and DAAC. DAAC involvement in naming also helps in making the data products discoverable and EOSDIS-unique. Data products must be assigned both a long name, which is meant to be human readable (and comprehensible), and a short name, which aids precise searching by keyword. The data product long name and short name are considered global attributes in that they are not associated with any particular variable but are related to the data product as a whole.

Data producers should seek a name that will be understandable to the target user community but also unique within EOSDIS. A reasonable data product name may already be in use (e.g., beginning with "MODIS/Aqua" or MODIS/Terra"), and care should be taken to avoid naming conflicts by consultation with the relevant data producers. Also consider interoperability when choosing names [26].

There is no universal file-naming convention for NASA Earth Science data products, apart from the DIWG recommendations regarding the components of file names provided in [26] (Rec. 3.8-3.11). However, filenames should be unique and human-readable and contain information that is descriptive of the contents.

4.2.1 Long Name

The data product long name (not to be confused with the CF `long_name` attribute for individual variables) is a scientific definition of the product. It should be as brief (but as complete) as possible, as it expands on the sometimes-cryptic corresponding short name. Specifically, the long name should be a character string from 45-90 characters to prevent wrapping and enhance readability and scannability in interfaces (such as Earthdata Search) [49]. The attribute LongName is synonymous with the attribute title used in netCDF documentation (see Appendix D.1).

The product long name should be included in the file as the global attribute LongName.

The recommended naming convention for data products includes the following information:

- The data source, usually the acronym/abbreviation for the project responsible for producing the data, but can also be the instrument (e.g., HIRDLS, the High Resolution Dynamics Limb Sounder), satellite (e.g., TRMM, the Tropical Rainfall Measuring Mission), or program (e.g., MEaSUREs, Making Earth System Data Records for Use in Research Environments); include both the instrument and satellite names to eliminate ambiguity (e.g., MODIS/Aqua).
- Science content (e.g., Aerosol, Precipitation).
- The general spatial distribution of the data (e.g., gridded, swath, orbit, *in situ*).
- For gridded data, the temporal resolution (e.g., daily, monthly).
- Processing level (e.g., L2 or Level 2).
- Spatial resolution (e.g., 3 km; if there are multiple resolutions in the product, then the highest resolution is stated).

Examples of data products that follow this naming convention are provided in

Table ***1***. Deviations from this format are sometimes necessary owing to characteristics of specific data products (Table 1, rows 7 and 8).

4.2.2 Short Name

EOSDIS developed the standard Earth Science Data Type (ESDT) [50] naming conventions to provide an efficient way to archive data products by name and for convenience in working with toolkits. The short name is an abbreviated version of the product name. A limit of 30 characters is recommended, where alphanumeric and "_" are the only acceptable characters. The restriction on the usage of spaces and special characters is to ensure compatibility with Earthdata Search and other search systems. The short name is included in the metadata as the global attribute ShortName and included in the data product's documentation. Data producers should contact the DAAC responsible for the data product's archiving to check if there are additional restrictions on short names.

Table 1. Examples of data products named using the recommended naming conventions.

Short Name	**Long Name**	**Comments**
OCO2_L2_Met	OCO-2 Level 2 meteorological parameters interpolated from global assimilation model for each sounding	Surface or 2-D [global; ECMWF model reanalysis data being collocated/interpolated unto OCO2 soundings (satellite-based) locations]
MYD09GQ	MODIS/Aqua Near Real Time (NRT) Surface Reflectance Daily L2G Global 250m SIN Grid	Surface or 2-D [global; satellite-based]
MOD05_L2	MODIS/Terra Total Precipitable Water Vapor 5-Min L2 Swath 1km and 5km - NRT	Surface or 2-D [vertically integrated quantity; global; satellite-based]
GPM_PRL1KU	GPM DPR Ku-band Received Power L1B 1.5 hours 5 km	Surface or 2-D [quasi-global; satellite-based]
GLDAS_NOAH10_M	GLDAS Noah Land Surface Model L4 Monthly 1.0 x 1.0 degree	2-D or Surface [global; model reanalysis]
SWDB_L3M10	SeaWiFS Deep Blue Aerosol Optical Depth and Angstrom Exponent Monthly Level 3 Data Gridded at 1.0 Degrees	Surface or 2-D [global; satellite-based]
AIRX2RET	Aqua AIRS Level 2 Standard Physical Retrieval (AIRS+AMSU)	3-D [global, satellite-based]
M2I3NPASM	MERRA2 inst3_3d_asm_Np: 3d, 3-Hourly, Instantaneous, Pressure-Level, Assimilation, Assimilated Meteorological Fields at 0.625 x 0.5 degree	3-D [global; MERRA-2 model reanalysis]

ATLAS_VEG_PLOTS_1541	Arctic Vegetation Plots ATLAS Project North Slope and Seward Peninsula, AK, 2998-2000	*in situ*
CARVE_L1_FTS_SPECTRA_1426	CARVE: L1 Spectral Radiance from Airborne FTS Alaska, 2012-2015	Airborne

4.3 Versions

The global attribute `product_version` is used to distinguish between versions of a given data product (e.g., produced using different processing algorithms or updates to calibration parameters). It is particularly important for users to know if they will acquire the latest version of a data product.

It is highly recommended to represent the data product version with an ordinal identifier (e.g., 1, 2, 3, etc.) that expresses its position in a series of data product publications. The data product version can be represented with both a major and minor version identifier (e.g., 2.1, 2.2, etc.). A minor version is used to identify selected files associated with a limited reprocessing of data (e.g., changes around a data anomaly that did not affect the rest of the data product). A change that affects the whole data product (e.g., complete reprocessing) would be considered a major version change. Guidance for setting version numbers should be sought from the DAAC hosting the data. Some guidance by the DIWG regarding version numbers can be found in [26] (Rec. 3.10). Whatever versioning scheme is used, it is understood that all files in a given data product were produced in a consistent manner.

Periodic reprocessing of data products can produce new versions with a distinct data product identifier. In general, data that are sufficiently different should be organized into separate data products. When data are reprocessed, the data producer must distinguish between major and minor version changes. Also, the nature of changes and the records to which they apply should be described for every version. In practice, a DAAC may choose to combine different minor versions of data into a single major version of a data product in the archive and only advance to the next major version upon reprocessing of the entire data product.

If possible, it is useful to use the same version for the data product as for the algorithm software used to generate the product (e.g., `PGEVersion`), to avoid confusing the data product users.

4.4 Representing Coordinates

4.4.1 Latitude and Longitude

Earth Science data files should be produced with complete information for all geospatial coordinates. This geolocation should be encoded in an interoperable way based on CF Conventions. Variables representing latitude and longitude must always explicitly include the `units` attribute; there is no default value. The recommended unit of latitude is `degrees_north` (valid range: -90 to 90 degrees) and unit of longitude is `degrees_east` (valid range: -180 to 180 degrees). Consider the spatial accuracy required to represent the position of data in the file when choosing the Latitude and Longitude parameter datatypes. Use FLOAT64 for Latitude and Longitude datatypes if precise geolocation of the data is required.

To support the widest range of software tools while avoiding storage of redundant geospatial coordinate data, practice the following guidelines that are detailed in [25] (Rec. 2) and [26] (Rec. 3):

- Specify coordinate boundaries by adding the `bounds` attribute [25] (Rec. 2.3).
    - For example, a producer can annotate the “time" coordinate with a `bounds` attribute with value “time_bounds.” The “time_bounds” variable would be a multi-dimensions array of the intervals for each value of variable “time.”
- Include horizontal and vertical (as necessary) attributes for data in grid structures.
    - For example, a producer can include attribute unit: `degrees_north` for the latitude coordinate variable; `degrees_east` for the longitude coordinate variable; and "meter" for a height coordinate variable.
- Store all coordinate data for a single file in coordinate variables only. No coordinate data, or any part thereof, should be stored in attributes, or as variable and group names [26] (Rec. 3.5).
- Files are required to contain the most applicable type of geospatial coordinates for the data. The decision whether to provide any additional types of geospatial coordinates is left to the data producer.
- The geolocation should be given as both grid mapping variable attributes and OGC Well-Known Text (WKT) whenever possible.

Note that different data user communities prefer different orders for including the latitude and longitude in the data files. Data producers should target the dominant user community for their products to decide upon the order, indicate clearly what the order is, and use self-describing formats.

4.4.2 Time

The CF Conventions represent time as an integer or float, with the `units` attribute set to the time unit since an epochal time, represented as YYYY-MM-DDThh:mm:ss (e.g., “seconds since 1993-01-01T00:00:00Z”). Unless there is a strongly justifiable reason not to do so, use the UTC Time Zone instead of alternative time zones.

The date-time information in the file should adhere to the following guidelines (detailed in [26], Rec. 3.11):

- Adopt the ISO 8601 standard [51] [52] for date-time information representation.
- If describing time intervals, the start time should appear before the end time.
- Date-time fields representing the temporal extent of a file’s data should appear before any other date-time field in the file name.
- All date-time fields should have the same format.

For gridded observations (Level 3 or Level 4), data can be aggregated using the time coordinate axis to record the different time steps (see Section 5). For example, this technique can be used to aggregate sub-daily observations (e.g., hourly, 4-hourly) into a single daily file.

When files contain only a single time slice, a time axis coordinate vector variable of length 1 should be included, so that the time information is easy to locate and the degenerate (i.e., one value of time for the entire file) time axis can improve performance when aggregating additional files over time ([25], Rec. 2.9; [26], Rec. 3.4).

4.4.3 Vertical

Data can be three-dimensional and, hence, include variables that describe a vertical axis. The most commonly used values to describe vertical coordinates are `level`, `pressure`, `height`, and `depth`. It is important to identify the vertical coordinates using the most common standard terminology, and including the following information is recommended:

- `long_name` This attribute can be something as simple as vertical level and can also be used to clarify the attribute `units`. The valid values are provided by the Unidata units library (UDUNITS) package [53] and include units of pressure, length, temperature, and density. For a dimensionless coordinate, it is acceptable to use the values `level`, `sigma_level`, or `layer`; however, it is recommended to add a `standard_name` attribute to describe the coordinate.
- `positive` This attribute refers to the direction of the increasing coordinate values, with a valid value of either `up` or `down`. If the vertical coordinate follows units of pressure, then this is not required. Variables representing dimensional height or depth axes must always explicitly include `units` and `positive` because there is no default value.

4.5 Data Quality

It is essential that users of scientific data products have access to complete (i.e., to the degree to which all knowledge is available at the time) and properly articulated (i.e., correctly described for the user to logically understand, discern, and make well-informed decisions) information about the data quality, including known issues and limitations. This information will help to inform users about the potential applications of the data products and prevent data misuse. Therefore, data products should include metadata pointing to documentation of the processes used for assessing data quality and supply the documentation to the DAACs for archiving and distribution. Data producers can work with the DAAC(s) and review boards to provide data quality information through an existing community-standardized format for describing quality (e.g., the data quality metadata model found in GHRSST [10]). For example, data quality information can be provided as a part of a data product users' guide.

The recommended contents for capturing and ensuring data quality are provided below. More detailed explanations on these, along with examples, can be found in the Data Management Plan Template for Data Producers [54]. In the discussion below, we use the term documentation to refer to somewhat extensive information that is typically stored separately from data files, with the metadata in the files including pointers (URLs) to such information.

4.5.1 Data Product Documentation

1. Document the process used, including data flows and organizations involved in assuring data quality. Provide reference to Interface Control Documents (ICDs), if any, between organizations

that have been or will be developed. If the ICD does not exist or is a work in progress, include the names and email addresses of the lead authors responsible for drafting the ICD. See [55] for an example of an ICD.

2. Calibration/Validation (Cal/Val – see Appendix B) is applicable only to missions in which Cal/Val is explicitly mandated. Provide documentation of the Cal/Val approach used, including sources of data, duration of the process, the targeted uncertainty budget that was used to assess performance, and the conditions under which Cal/Val are conducted. As Cal/Val data sources change or are re-processed, ensure that the information is kept up to date in a publicly accessible location with reference to the relevant geospatial and temporal coverage information that is directly applicable to those Cal/Val data products.
3. Provide a description of how quality flags or indicators (see Appendix B) are used in the product and explain their meanings. The following are general considerations regarding quality flags and indicators:
    a. Define and create indicators to represent quality of a data product from different aspects (e.g., data dropout rate of a sea surface temperature data product).
    b. Ensure that quality flags are related to a quantifiable metric that directly relates to the usefulness, validity, and suitability of the data.
    c. Identify quantifiable data quality criteria, such as confidence levels and the values of quality flags, which can be used as criteria for refining search queries. Provide quality and measurement state information in CF-compliant attributes [22]: `flag_values`, `flag_masks`, and `flag_meanings` [22] (Section 3.5). The choice of `flag_values` vs. `flag_mask` depends on the use case. The `flag_values` and `flag_masks` may or may not be used together. An example of a complex case in which they can both be used is illustrated in [22], Section 1.7, example 3.5. In all cases, `flag_meanings` is used in conjunction with either `flag_masks` or `flag_values`.
    d. Provide ancillary quality and uncertainty flags to facilitate detection of areas that are likely to contain spurious data (e.g., ice in unexpected places).
    e. Provide pixel-level (or measurement-level) uncertainty information where possible and meaningful. Provide the confidence level (e.g., 95%) to indicate the statistical significance.
    f. Provide data quality variables and metadata along with detailed documentation on how the metadata are derived and suggestions on how to interpret them or use them in different applications.
    g. Provide definition and description of each data quality indicator, including the algorithms and data products used to derive the quality information and description of how each quality indicator can be used.
    h. Provide examples of idealized quality flag/indicator combinations that would likely yield optimal quality filtering (i.e., minimized bias, uncertainty, and spurious observations) for science in a particular domain of research.
4. While a user should be able to independently derive and extract quality summary information from the data files (i.e., via quality flags and quality indicators), the quality summary should also

be documented and disseminated at the time that a new dataset version is published. The quality summary should at least be a high level overview of strengths and limitations of the dataset, and should be directly traceable and reproducible by the variables within the dataset, such as by referencing the quality flags and indicators used to derive the summary. For example, the quality summary may describe the overall percentage of data that is either missing from the dataset (due to pre-processing QA/QC) or that may be optionally discarded (at the discretion of the data user) due to quality conditions that are expressed by the quality flags and indicators.

5. Provide documentation of methods of estimating uncertainty and how they are included in the data product. Provide documentation of known issues and caveats for users and consider leveraging DAAC resources for more expedient updating and publication of this information (e.g., forums, Web announcements). Also include citations and references to the data used in the validation process.
6. Provide quality summary information such as attributes describing the percent of observations that are missing or are in each quality category.

4.5.2 File Metadata

1. Include the uncertainties in the delivered data, with the level of detail dependent on the size of the uncertainty information. For example, these can be expressed per data value, per file, or at the data product level.
2. Provide pointers to the ancillary data products that are used for quality assessments, Cal/Val, uncertainty budget validation, uncertainty quantification, and uncertainty characterization.
3. Implement quality flags and indicators consistent with the documentation discussed above (Item #3 in section 4.5.1).

Ensure compliance with metadata standards related to data quality – International Organization for Standardization (ISO) 19157 [56], CF Conventions [22] including those for flags and indicators, ACDD [57], and ISO 8601 [51] [52]. Plan on using an automated compliance checker.

4.6 Global Attributes

Global attributes (i.e., those that apply to an entire file rather than to a particular group or variable) improve data discoverability, documentation, and usability. Descriptions of the recommended global attributes, according to CF, Attribute Conventions for Data Discovery (ACDD) [57], and other conventions, can be found in Appendix D.

4.6.1 Provenance

Data provenance consists of the origins, lineage, custody, and ownership of data, and must be included in metadata for transparency and reproducibility. File-level provenance can be combined with the data-product-level provenance to help the user ascertain the overall data product provenance. When describing provenance, include information about the environment used to create the data product (e.g., software version, processing system, processing organization) and the context of the run (e.g., production time, list of input data, and ancillary files).

The recommended provenance metadata for the processing environment include the attributes: `AlgorithmType`, `AlgorithmVersion`, `history`, `ProcessingCenter`, `PGEVersion`, `PGE_Name`, and `ProcessingEnvironment`. The recommended provenance metadata for the run context are: `InputPointer`, `PGE_EndTime`, `PGE_StartTime`, `ProductionTime`, `RangeBeginningDate`, `RangeBeginningTime`, `RangeEndingDate`, `RangeEndingTime`, and `VersionId`.

4.7 Variable Attributes

Variables should also contain specific attributes that describe the data within each file. The recommended variable attributes given by the CF and the ACDD conventions can be found in Appendix E. See also the CF Metadata Template [58].

5 DATA COMPRESSION, CHUNKING AND PACKING

Data compression and chunking are two storage features provided by the HDF5 library and available through the netCDF-4 API. HDF5's internal compression can reduce the space taken up by variables, especially those with many fill values or value repetition. The saved space can pay significant dividends in both storage space and transmission speed over the network. HDF5 includes a compression method referred to as "deflation", based on the common compressor gzip (itself based on the Lempel-Ziv algorithm). Deflation levels run from 1 to 9, with storage efficiency and time to compress increasing with each level. A level of 5 is often a good compromise. HDF5/NetCDF-4 variables are individually compressed within the file, which means that applications only need to uncompress those variables of interest, and not the whole file, as would be necessary for external compression methods such as gzip or bzip. HDF5/NetCDF-4 variables are also "chunked," which means that each variable is stored as a sequence of separate chunks in the file. If compression is used, each chunk is compressed separately. This allows read programs to decompress just the chunks needed for a read request, and not the entire variable, resulting in even greater IO efficiencies. Chunking also can allow a calling program to retrieve segments of data more efficiently when those data are stored in Object Storage (see Appendix B, Glossary). Note that the DIWG

recommends using only the DEFLATE compression filter on NetCDF-4 and NetCDF-4-Compatible HDF5 Data. Also, applying the netCDF-4 shuffle filter before deflation can significantly improve the data compression ratio for multidimensional netCDF-4 variables.

An alternative way to reduce data size is to apply a scale and offset factor to the data values, allowing the values to be saved as a smaller datatype, such as a 16-byte short int. This technique, known as "packing," is appropriate for data with a limited dynamic range. The attributes needed to reconstruct the original values are:

- scale_factor
- add_offset

The equation to reconstruct the original values is:

final_data_value = scale_factor * packed_data_value + add_offset

The values for scale_factor and add_offset may be selected by the data producer, or automatically computed to minimize storage size by a packing utility such as **ncpdq** (part of the NCO package [59]). The DIWG recommends that packing be employed only when data are stored as integers [25] (Rec. 2.6).

An additional benefit of packing is that it can sometimes make the data more compressible via deflation. That is, packing followed by the netCDF-4 shuffle filter followed by deflation can result in very significant data compression.

Chunking is appropriate for variables with a large number of values, particularly multidimensional ones. It is helpful to consider the most likely pattern of data usage. However, where this is unknown or widely varied, "balanced chunking" is recommended, i.e., balance access speeds for time-series and geographic cross-sections, the two most-common end member geometries of data access. For example, Unidata has an algorithm for balanced chunking [60]. The DIWG recommends using balanced chunking for multidimensional variables contained in grid structures [25] (Rec. 2.11).

The HDF5 API supports the ability to save data in chunked and/or compressed form at the time of creation, but this requires writing the whole variable at once. In addition, the following command-line utilities can be used to chunk and compress files after they have been written:

- h5repack (part of the HDF5 library).
- nccopy (part of the netCDF library).
- ncks (part of the NCO package [61]).

These utilities are also useful in experimenting with different compression levels and chunking schemes.

6 TOOLS FOR DATA PRODUCT TESTING

The following steps should be followed to test compliance and usability of a new data product.

1. Perform a data inspection using a data dump (i.e., displaying the data) to check for successful ingest of the metadata, that global attributes appear as expected in the relevant search systems, and that data agrees with the product user guide (Section 6.1).

2. Automated compliance checkers should be used to test the product against format standards (Section 6.2).
3. If problems are found, some data producers may find it useful to inspect and edit the metadata using tools described in Section 6.3.
4. Once all the necessary changes are known, the data production code should be modified accordingly.
5. The product should be tested with tools that will likely be used on the product (Section 6.4).
6. Producer may wish to validate the packaging decisions result in the desired size / performance trade-off.

6.1 Data Inspection

Dumping the data and inspecting them is a useful first check or troubleshooting technique. It can reveal obvious problems with standards compliance and consistency with the data product users' guide. Useful tools for data inspection of netCDF-4 and HDF5 files are summarized in Table 2 (but also see [62]).

Table 2. Useful tools for inspecting NetCDF-4 and HDF5 data. The types of tools include command-line interface (CLI) and desktop graphical user interface (GUI).

Tool	Type	Access	Capabilities
HDFView	GUI	Website [63]	Read HDF5 and netCDF-4 files; views any data object; select "Table" from menu bar and then "export to text" or "export to binary."
Panoply	GUI	Website [64]	Read HDF5 and netCDF-4 files; the Array tab displays the actual data values that can further be edited in a spreadsheet.
h5dump	CLI	HDF5 library [20]; Anaconda [65]	Read and dump HDF5 and netCDF-4 files.
IDL	CLI	IDL [66]	Interactive Data Language. provides built-in support for several data sources, data types, file formats, and file sizes
IDV	CLI	IDV [67]	Integrated Data Viewer. 3D geoscience visualization and analysis tool that gives users the ability to view and analyze geoscience data in an integrated fashion.
ncdump	CLI	NetCDF-4 C library [68]	Dump netCDF-4 content to ASCII format.
ncks	CLI	NCO Toolkit [69]; Anaconda [65] add-on	Read and dump HDF5 and netCDF-4 files.
NCL	CLI	NCAR Command Language [70]	Interpreted language designed for scientific data analysis and visualization[2]

6.2 Compliance Checkers

Compliance checkers should be used while data products are being developed to ensure that the metadata fields are all populated and are meaningful. The following are recommended compliance checkers:

- HPD [21] is a design tool as indicated in Section 3.1, which can be used for checking which metadata in a given file are CF- and ACDD-compliant and which are not.
- Metadata Compliance Checker is a Web-based tool and service designed by the Physical Oceanography DAAC for netCDF and HDF formats [71].
- CF-Convention Compliance Checker was developed by Hadley Centre for Climate Prediction and Research for netCDF formats [72].
- CFChecker developed by Decker [73].
- Dismember developed by NCO [74].

[2] NCL was put into maintenance mode in 2019.

- Integrated Ocean Observing System (IOOS) Compliance Checker [75]

6.3 Internal Metadata Editors

Data editors can be useful in tweaking metadata internal to the data files when debugging samples that surface problems during testing. Once the metadata (or data) have been corrected, of course, the data processing code typically needs to be modified for the actual production runs.

Several useful tools are available for editing data in netCDF-4 and HDF5 formats are summarized in Table 3 (but also see [62]).

Table 3. Useful tools for editing NetCDF-4 and HDF5 metadata and data. The types of tools include command-line interface (CLI) and desktop graphical user interface (GUI).

Tool	Type	Access	Capabilities
HPD	GUI	Website [21]	Add or remove metadata to make the file CF- and/or ACDD-compliant; include ISO Metadata in an Earth Science data file.
HDFView	GUI	Website [63]	Create, edit, and delete content of netCDF-4 and HDF5 files.
ncatted	CLI	NCO Toolkit [69]; Anaconda [65] add-on	Edit netCDF-4 global, group and variable attributes.
ncks	CLI	NCO Toolkit [69]; Anaconda [65] add-on	For netCDF-4: subset, chunk, compress, convert between versions, copy variables from one file to another, merge files, print.
Ncrename	CLI	NCO Toolkit [69]; Anaconda [65] add-on	Rename groups, dimensions, variables and attributes of netCDF-4 files.
ncdump	CLI	NetCDF-4 C library [68]	Print the internal metadata of netCDF-4 files.
ncgen	CLI	NetCDF-4 C library [68]	Convert ASCII files to netCDF-4 format.

6.4 End-User Tools

Two generalized GUI tools in particular work with a large variety of HDF (and netCDF) products: Panoply [64] and HDFView [63]. Thus, these are recommended for at least minimal testing of data products before their release to users. If a data product cannot be read (and ideally plotted) through these tools, it indicates serious problems in the product. Appendix C provides illustrations of these tools. In addition, it is helpful to test with tools that are in wide use by the target community for a data product, such as GIS tools for land processes products. In the yearly EOSDIS user survey for 2018, the tools mentioned by more than 10 users are included the following, in descending order of mentions:

Table 4: Frequently used end-user tools

Tool	Source	URL
SNAP (Sentinel Application Platform)	European Space Agency	https://step.esa.int/main/toolboxes/snap/
Python	(many)	https://www.earthdatascience.org/courses/intro-to-earth-data-science/python-code-fundamentals/use-python-packages/
QGIS	QGIS.org	https://qgis.org/en/site/
saga gis	Saga-GIS.org	https://www.saga-gis.org
GDAL (Geospatial Data Abstraction Library)	OSGeo	https://gdal.org/
ArcGIS	ESRI	http://desktop.arcgis.com/en/
R	R Project for Statistical Computing	https://www.r-project.org/
MATLAB	Mathworks	https://www.mathworks.com/products/matlab.html
Octave (similar to	GNU.org	https://www.gnu.org/software/octave/

Tool	Source	URL
MATLAB)		
Google Earth Engine	Google	https://earthengine.google.com/
ERDAS IMAGINE	Hexagon Geospatial	https://www.hexagongeospatial.com/products/power-portfolio/erdas-imagine

7 DATA PRODUCT DIGITAL OBJECT IDENTIFIERS

A Digital Object Identifier (DOI) is a unique alphanumeric character string (i.e., handle) used to identify an object. A DOI is permanent, such that when it is registered, it can be used to locate the object to which it refers permanently. Since their introduction in 2000, DOIs have been routinely assigned to journal articles and cited by the scientific community. Use of DOIs for data products, however, is more recent but equally important for universal referencing and discoverability of data, as well as for proper attribution and citation.

The DOI handle is composed of a prefix that includes characters to identify the registrant and a suffix that includes the identification number of the registered object. In addition to the DOI handle, a Web address is assigned by the DOI registration service provider. For a data product, its DOI typically leads to a Web landing page (for guidelines, see [76] [77]) that provides information about the data product and services for users to obtain the data. One of the key benefits of assigning a DOI to a data product is that even if the Web address changes, the DOI remains valid. This means that a DAAC can change the Web address of a data product without affecting the validity of references made in published literature. In addition, the data publisher could change, but the DOI is unaffected. For a detailed description of DOI, see the DOI Handbook [78]. The ESDIS Project has established procedures for managing DOIs for EOSDIS data [79]. The format of the DOIs managed by the ESDIS Project is 10.5067/*[suffix].* Here the suffix uniquely identifies the object. It can be structured (containing meaningful information about the digital object) or opaque (any combination of alphanumeric characters, usually generated randomly and not having any semantic content).

Data producers should work with the DAACs to assign DOIs to their data products. The requests for DOI registration are made to the ESDIS Project by a DAAC.

The ESDIS Project uses a two-step process for registering DOIs. First, DOIs are reserved, so that data producers can start using them in the metadata while generating the products. Information about the DOI should be included in the data product metadata. In particular, the DOI resolving authority and the DOI identifier must be included as global attributes (see [80]). When a data product is ready to be delivered to the DAAC for public release, the DOI is registered. Until the DOI is registered, it can be modified or deleted (withdrawn). However, once registered, the DOI becomes permanent.

8 PRODUCT DELIVERY AND PUBLICATION

The responsibilities for generating data products and making them available to the user community are shared between data producers and the DAACs. The details of the processes leading to data delivery and publication vary depending on the type of data producer as well as the DAAC. The general process for adding new data to EOSDIS as well as the requirements and responsibilities of data producers and DAACs are shown in [81]. Each of the twelve EOSDIS DAACs have established publication workflows that account for the heterogeneous suite of missions, instruments, data

producers, data formats, and data services managed by EOSDIS. While some aspects of data publication vary across DAACs, the primary phases of the data publication processes are generally the same: obtain the data and related information from data producers, describe the data with metadata and documentation[3], and release the data for access by the user community. The specifics of data delivery and publication, such as schedules, interfaces, workflow and procedures for submitting data product updates, are best established by communications between the data producers and their assigned DAACs.

9 REFERENCES

[1] ESDIS, "Earth Science Data Systems (ESDS) Program," 3 February 2020. [Online]. Available: https://earthdata.nasa.gov/about. [Accessed 24 April 2020].

[2] ESDIS, "Earth Science Data and Information System (ESDIS) Project," 16 January 2020. [Online]. Available: https://earthdata.nasa.gov/esdis. [Accessed 24 April 2020].

[3] ESDIS, "Earth Science Data System Working Groups," 25 February 2020. [Online]. Available: https://earthdata.nasa.gov/collaborate/esdswg. [Accessed 24 April 2020].

[4] ESDS Program, "Data Processing Levels," 23 August 2019. [Online]. Available: https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels. [Accessed 24 April 2020].

[5] GES DISC, "GES DISC Data and Metadata Recommendations to Data Providers," 4 November 2017. [Online]. Available: https://docserver.gesdisc.eosdis.nasa.gov/public/project/DataPub/GES_DISC_metadata_and_data_formats.pdf. [Accessed 24 April 2020].

[6] SEDAC, "Documentation for the India Village-Level Geospactial Socio-Economic Data Set, v1 (1991, 2001) - usage of the SEDAC Data Documentation Template," [Online]. Available: https://doi.org/10.7927/H43776SR. [Accessed 24 April 2020].

[7] PO DAAC, "PO.DAAC data management best practices," [Online]. Available: https://podaac.jpl.nasa.gov/PO.DAAC_DataManagementPractices. [Accessed 24 April 2020].

[8] NSIDC DAAC, "NSIDC Page for SNOWEX Data Providers," 2020. [Online]. Available: https://nsidc.org/data/snowex/information-snowex-data-providers. [Accessed 24 April 2020].

[3] Data documentation should include at least a User's Guide. If the product is a geophysical retrieval, then providing an Algorithm Theoretical Basis Document is recommended as well.

[9] ORNL DAAC, "ORNL DAAC detailed submission guidelines," [Online]. Available: https://daac.ornl.gov/submit/. [Accessed 24 April 2020].

[10] GHRSST Science Team, "The Recommended GHRSST Data Specification (GDS) 2.0, Document Revision 5," GHRSST International Project Office, 2010.

[11] ESDIS, "ESDIS Standards Office (ESO)," 27 February 2020. [Online]. Available: https://earthdata.nasa.gov/esdis/eso. [Accessed 24 April 2020].

[12] ESDIS, "Glossary," 28 January 2020. [Online]. Available: https://earthdata.nasa.gov/learn/user-resources/glossary. [Accessed 24 April 2020].

[13] ESDIS Project, "Common Metadata Repository," 12 March 2020. [Online]. Available: https://earthdata.nasa.gov/eosdis/science-system-description/eosdis-components/cmr. [Accessed 24 April 2020].

[14] G. Asrar and H. K. Ramapriyan, "Data and Information System for Mission to Planet Earth," *Remote Sensing Reviews,* vol. 13, pp. 1-25. https://doi.org/10.1080/02757259509532294, 1995.

[15] H. K. Ramapriyan and J. Moses, "NASA Earth Science Data Preservation Content Specification," 5 August 2019. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references/preservation-content-spec. [Accessed 25 June 2020].

[16] H. Ramapriyan, G. Peng, D. Moroni and C.-L. Shie, "Ensuring and Improving Information Quality for Earth Science Data and Products," *D-Lib Magazine,* vol. 23, no. July/August 2017 Number7/8, 2017.

[17] ESO, "Standards and Requirements," ESDIS Project, 18 February 2020. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references. [Accessed 24 April 2020].

[18] ESO, "Standards and Practices," ESDIS Project, 18 February 2020. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references#standards-and-practices. [Accessed 23 April 2020].

[19] ESO, "netCDF-4/HDF5 File Format," ESDIS Project, 5 August 2019. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references/netcdf-4hdf5-file-format. [Accessed 24 April 2020].

[20] ESO, "HDF5 Data Model, File Format and Library – HDF5 1.6," ESDIS Project, 5 August 2019. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references/hdf5. [Accessed 24 April 2020].

[21] A. Jelenak, "HDF Product Designer," 11 February 2019. [Online]. Available: https://wiki.earthdata.nasa.gov/display/HPD/HDF+Product+Designer. [Accessed 24 April 2020].

[22] ESDIS, "Climate and Forecast (CF) Metadata Conventions," ESDIS Standards Office, 5 August 2019. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references/climate-and-forecast-cf-metadata-conventions. [Accessed 24 April 2020].

[23] B. Eaton and et al., "NetCDF Climate and Forecast (CF) Metadata Conventions (section 2.7)," 24 April 2020. [Online]. Available: http://cfconventions.org/cf-conventions/cf-conventions.html#groups. [Accessed 24 April 2020].

[24] "Encoding of Swath Data in the Climate and Forecast Convention," 19 June 2018. [Online]. Available: https://github.com/Unidata/EC-netCDF-CF/blob/master/swath/swath.adoc. [Accessed 24 April 2020].

[25] DIWG, "Dataset Interoperability Recommendations for Earth Science, ESDS-RFC-028v1.3," ESDIS Project, 12 November 2019. [Online]. Available: https://earthdata.nasa.gov/user-resources/standards-and-references/dataset-interoperability-recommendations-for-earth-science. [Accessed 24 April 2020].

[26] DIWG, "Dataset Interoperability Recommendations for Earth Science: Part 2, 423-ESO-036," 12 November 2019. [Online]. Available: https://wiki.earthdata.nasa.gov/display/ESO/ESO+Review+-+Dataset+Interoperability+Working+Group+Recommendations+-+Part+2. [Accessed 24 April 2020].

[27] "NCEI NetCDF Templates v2.0," 16 April 2018. [Online]. Available: https://www.nodc.noaa.gov/data/formats/netcdf/v2.0/. [Accessed 24 April 2020].

[28] ESDIS Project, "Earthdata Search," [Online]. Available: https://search.earthdata.nasa.gov/search. [Accessed 27 April 2020].

[29] Adobe, "TIFF," 2019. [Online]. Available: https://www.adobe.io/open/standards/TIFF.html. [Accessed 24 April 2020].

[30] wikipedia, "Geographic information system," wikipedia, 16 April 2020. [Online]. Available: https://en.wikipedia.org/wiki/Geographic_information_system. [Accessed 24 April 2020].

[31] OGC, "Welcome to the Open Geospatial Consortium," 2020. [Online]. Available: https://www.opengeospatial.org. [Accessed 24 April 2020].

[32] ESO, "GeoTIFF File Format, ESDS-RFC-040v1.1," 16 September 2019. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references/geotiff. [Accessed 24 April

2020].

[33] COG, "Cloud Optimized GeoTIFF: An imagery format for cloud-native geospatial processing," [Online]. Available: https://www.cogeo.org/. [Accessed 24 April 2020].

[34] ESO, "Instructions to RFC Authors," ESDIS Project, 5 August 2019. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references/instructions-to-rfc-authors. [Accessed 24 April 2020].

[35] E. Northup, G. Chen, K. Aikin and C. Webster, "ICARTT File Format Standards V2.0," January 2017. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references/icartt-file-format. [Accessed 24 April 2020].

[36] ESDIS Standards Office, "ASCII File Format Guidelines for Earth Science Data, ESDS-RFC-027v1.1," May 2016. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references/ascii-file-format-guidelines-for-earth-science-data. [Accessed 24 April 2020].

[37] OGC, "GeoPackage: An Open Format for Geospatial Information," OGC, 2018. [Online]. Available: https://www.geopackage.org/. [Accessed 24 April 2020].

[38] W3C, "Extensible Markup Language (XML)," 11 October 2016. [Online]. Available: https://www.w3.org/XML/. [Accessed 24 April 2020].

[39] ESO, "OGC KML," ESDIS Standards Office, 5 August 2019. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references/ogc-kml. [Accessed 24 April 2020].

[40] Esri, "Shapefiles," Esri, [Online]. Available: https://doc.arcgis.com/en/arcgis-online/reference/shapefiles.htm. [Accessed 24 April 2020].

[41] Wikipedia, "Shapefile," Wikipedia, 6 March 2020. [Online]. Available: https://en.wikipedia.org/wiki/Shapefile. [Accessed 24 April 2020].

[42] Esri, "Geoprocessing considerations for shapefile output," 24 April 2009. [Online]. Available: http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Geoprocessing%20considerations%20for%20shapefile%20output. [Accessed 24 April 2020].

[43] ESO, "HDF-EOS5 Data Model, File Format and Library," ESDIS Standards Office, 5 August 2019. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references/hdf-eos5. [Accessed 24 April 2020].

[44] A. Taaheri and K. Rodrigues, "HDF-EOS5 Data Model, File Format and Library," May 2016. [Online]. Available: https://cdn.earthdata.nasa.gov/conduit/upload/4880/ESDS-RFC-008-v1.1.pdf. [Accessed 24 April 2020].

[45] T. Stevens, "GCMD Keyword Access," 9 April 2019. [Online]. Available: https://wiki.earthdata.nasa.gov/display/CMR/GCMD+Keyword+Access. [Accessed 24 April 2020].

[46] ESDIS, "Index of /static/kms," 2019. [Online]. Available: https://gcmdservices.gsfc.nasa.gov/static/kms/. [Accessed 27 April 2020].

[47] ESDIS, "NASA's GCMD releases the Keyword Governance and Community Guide Document, Version 1.0," 3 March 2020. [Online]. Available: https://earthdata.nasa.gov/learn/articles/nasa-s-gcmd-releases-the-keyword-governance-and-community-guide-document-version-1-0. [Accessed 27 April 2020].

[48] CEOS, "CEOS Analysis Ready Data," Committee on Earth Observation Satellites, [Online]. Available: http://ceos.org/ard/. [Accessed 27 April 2020].

[49] Butterick, "Responsive web design," [Online]. Available: https://practicaltypography.com/responsive-web-design.html. [Accessed 27 April 2020].

[50] ESDIS, "EOSDIS Glossary," 28 January 2020. [Online]. Available: https://earthdata.nasa.gov/learn/user-resources/glossary#ed-glossary-e. [Accessed 27 April 2020].

[51] ISO, "ISO 8601-1:2019 Date and time -- Representations for information interchange -- Part 1: Basic rules," February 2019. [Online]. Available: https://www.iso.org/standard/70907.html. [Accessed 27 April 2020].

[52] ISO, "ISO 8601-2:2019 Date and time -- Representations for information interchange -- Part 2: Extensions," February 2019. [Online]. Available: https://www.iso.org/standard/70908.html. [Accessed 27 April 2020].

[53] Unidata, "UDUNITS 2.2.26 Manual," 2020. [Online]. Available: https://www.unidata.ucar.edu/software/udunits/udunits-current/doc/udunits/udunits2.html. [Accessed 27 April 2020].

[54] Earth Science Data Systems (ESDS) Program, HQ SMD, "Data Management Plan (DMP) Template for Data Producers," 31 January 2019. [Online]. Available: https://wiki.earthdata.nasa.gov/download/attachments/118138197/ESDIS05161_DMP%20for%20DPs%20template%20-%20V1_20190131.pdf. [Accessed 27 April 2020].

[55] ESDIS Project, "ICD Between the ICESat-2 Science Investigator-led Processing System (SIPS) and the National Snow and Ice Data Center (NSIDC) Distributed Active Archive Center (DAAC) - 423-ICD-007, Revision A," NASA GSFC, Greenbelt, MD, 2016.

[56] ISO, "ISO 19157:2013 Geographic Information - Data Quality," December 2013. [Online]. Available: https://www.iso.org/standard/32575.html. [Accessed 27 April 2020].

[57] ESIP Documentation Cluster, "Attribute Conventions for Data Discovery," 24 January 2017. [Online]. Available: http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery. [Accessed 27 April 2020].

[58] D. Auty, "Template for CF (Climate-Forecast) Compliance," 31 December 2018. [Online]. Available: https://wiki.earthdata.nasa.gov/display/ESDSWG/Template+for+CF+(Climate-Forecast)+Compliance. [Accessed 27 April 2020].

[59] NCO, "NCO 4.9.3-alpha05 User Guide," 16 April 2020. [Online]. Available: http://nco.sourceforge.net/nco.html#ncpdq-netCDF-Permute-Dimensions-Quickly. [Accessed 24 April 2020].

[60] Developers@Unidata, "Chunking Data: Choosing Shapes," 2019. [Online]. Available: https://www.unidata.ucar.edu/blogs/developer/en/entry/chunking_data_choosing_shapes. [Accessed 13 August 2019].

[61] NCO, "NCO 4.9.3-alpha06 User Guide," 16 April 2020. [Online]. Available: http://nco.sourceforge.net/nco.html#Chunking. [Accessed 24 April 2020].

[62] ESDIS, "Data Tools," 16 January 2020. [Online]. Available: https://earthdata.nasa.gov/earth-observation-data/tools. [Accessed 27 April 2020].

[63] The HDF Group, "HDF View," 2006. [Online]. Available: https://www.hdfgroup.org/downloads/hdfview/. [Accessed 24 April 2020].

[64] NASA, "Panoply netCDF, HDF and GRIB Data Viewer," 28 February 2020. [Online]. Available: https://www.giss.nasa.gov/tools/panoply/. [Accessed 24 April 2020].

[65] Anaconda, "Your data science toolkit," [Online]. Available: https://www.anaconda.com/products/individual. [Accessed 27 April 2020].

[66] Harris Geospatial Solutions, "IDL: Exract Meaningful Visualizations from Complex Numerical Data," 2019. [Online]. Available: https://www.harrisgeospatial.com/Software-Technology/IDL. [Accessed 28 March 2019].

[67] UCAR, "Integrated Data Viewer," [Online]. Available: https://www.unidata.ucar.edu/software/idv/. [Accessed 17 March 2019].

[68] UCAR, "Unidata Data Services and Tools for Geoscience," 2020. [Online]. Available: https://www.unidata.ucar.edu/downloads/netcdf/. [Accessed 27 April 2020].

[69] NCO, "Bienvenue sur le netCDF Operator (NCO) site," 14 February 2020. [Online]. Available: http://nco.sourceforge.net/. [Accessed 24 April 2020].

[70] NCAR, "The NCAR Command Language (Version 6.4.0) [Software]. (2017)," Boulder, Colorado: UCAR/NCAR/CISL/VETS, [Online]. Available: http://dx.doi.org/10.5065/D6WD3XH5. [Accessed 2 August 2019].

[71] PO DAAC, "Metadata Compliance Checker," [Online]. Available: https://podaac-tools.jpl.nasa.gov/mcc/. [Accessed 27 April 2020].

[72] Hadley Centre for Climate Prediction and Research, "CF-Convention Compliance Checker for NetCDF Format," May 2019. [Online]. Available: http://puma.nerc.ac.uk/cgi-bin/cf-checker.pl. [Accessed 24 April 2020].

[73] M. Decker, "CFchecker - a CF-1.x compliance checker," [Online]. Available: https://bitbucket.org/mde_/cfchecker. [Accessed 27 April 2020].

[74] NCO, "Dismembering Files," 16 April 2020. [Online]. Available: http://nco.sourceforge.net/nco.html#ncdismember. [Accessed 24 April 2020].

[75] IOOS, "IOOS Compliance Checker," [Online]. Available: https://compliance.ioos.us/index.html. [Accessed 24 June 2020].

[76] ESDIS Project, "DOI Landing Page," 13 October 2016. [Online]. Available: https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS/DOI+Landing+Page. [Accessed 27 April 2020].

[77] ESDIS Project, "DOI Documents," 26 March 2020. [Online]. Available: https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS/DOI+Documents. [Accessed 27 April 2020].

[78] International DOI Foundation, "DOI Handbook," 19 December 2019. [Online]. Available: http://www.doi.org/hb.html. [Accessed 27 April 2020].

[79] ESDIS Project, "Digital Object Identifiers for EOSDIS," 3 October 2019. [Online]. Available: https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS. [Accessed 27 April 2020].

[80] ESDIS Project, "DOI Background Information," 27 April 2020. [Online]. Available: https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS/DOI+Background+Information. [Accessed 27 April 2020].

[81] ESDS Program, "Adding New Data to EOSDIS," 1 October 2019. [Online]. Available: https://earthdata.nasa.gov/collaborate/new-missions. [Accessed 27 April 2020].

[82] PO.DAAC, "Data Best Practices - Metadata Conventions," [Online]. Available: https://podaac.jpl.nasa.gov/PO.DAAC_DataManagementPractices#Metadata%20Conventions. [Accessed 27 April 2020].

[83] C. Davidson and E. Masuoka, "VIIRS Science Software Delivery Guide," LAADS DAAC, NASA GSFC, Greenbelt, MD, 2019.

[84] M. D. Wilkinson, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data,* vol. 3, 15 March 2016.

[85] ESDIS, "Metadata Requirements – Base Reference for NASA Earth Science Data Products, 423-RQMT-003, Revision B," 18 December 2018. [Online]. Available: https://ops1-cm.ems.eosdis.nasa.gov/cm2/flow/docDirectory/mCdl?execution=e2s1. [Accessed 24 April 2020].

[86] "Uniform Resource Identifier (URI) Schemes," 13 April 2020. [Online]. Available: https://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml. [Accessed 27 April 2020].

[87] OGC, "Simple Feature Access - Part 1: Common Architecture," OGC, 28 May 2011. [Online]. Available: https://www.opengeospatial.org/standards/sfa. [Accessed 27 April 2020].

[88] "Reverse DNS Look-up," [Online]. Available: https://remote.12dt.com/. [Accessed 27 April 2020].

[89] NOAA NGDC, "LI Lineage," NOAA NGDC, 27 November 2019. [Online]. Available: https://www.ngdc.noaa.gov/wiki/index.php/LI_Lineage. [Accessed 27 April 2020].

[90] NOAA EDM, "ISO 19115 and 19115-2 CodeList Dictionaries," ESIP, 3 October 2018. [Online]. Available: http://wiki.esipfed.org/index.php/ISO_19115_and_19115-2_CodeList_Dictionaries. [Accessed 27 April 2020].

10 AUTHORS' ADDRESSES

Hampapuram K. Ramapriyan
Telephone: 805.402.7125
Email: Hampapuram.Ramapriyan@ssaihq.com

Peter J. T. Leonard
Telephone: 301.352.4659
Email: pleonard@sesda.com

11 CONTRIBUTORS AND EDITORS

(Editors' names are shown in **bold**.)

Ed Armstrong, JPL/CalTech

Walter E. Baksin, SSAI & NASA/LaRC

Jeanne Beatty, ADNET & NASA/GSFC

Robert R. Downs, SEDAC

George Huffman, NASA/GSFC

Aleksandar Jelenak, The HDF Group

Siri Jodha Khalsa, NSIDC DAAC

Amanda Leon, NSIDC DAAC

Peter J.T. Leonard, ADNET & NASA/GSFC

Wen-Hao Li, JPL/CalTech

Christopher Lynnes, NASA/GSFC

David Moroni, JPL/CalTech

John Moses, NASA/GSFC

Dana Ostrenga, ADNET & NASA/GSFC

Hampapuram K. Ramapriyan, SSAI & NASA/GSFC

Justin Rice, NASA/GSFC

Elliot Sherman, SSAI & NASA/GSFC

Tammy Walker, ORNL DAAC

Lalit Wanchoo, ADNET & NASA/GSFC

Yaxing Wei, ORNL DAAC

Jessica N. Welch, ORNL DAAC

Robert Wolfe, NASA/GSFC

APPENDIX A. ABBREVIATIONS AND ACRONYMS

ACDD	Attribute Convention for Data Discovery
AIRS	Atmospheric Infrared Sounder (on Aqua)
AMSU	Advanced Microwave Sounding Unit (on Aqua)
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
ATBD	Algorithm Theoretical Basis Document
Cal/Val	Calibration/Validation
CF	Climate and Forecast Metadata Conventions
ChEBI	Chemical Entities of Biological Interest
CLI	Command Line Interface
CMR	Common Metadata Repository
CRS	Coordinate Reference System
CSDMS	Community Surface Dynamics Modelling System
CSV	Comma-Separated Variables
DAAC	Distributed Active Archive Center
DIWG	Dataset Interoperability Working Group
DOI	Digital Object Identifier
DPR	Dual-Frequency Precipitation Radar (on GPM)
EASE Grid	Equal Area Scalable Earth Grid
ENVO	Environment Ontology
EOS	Earth Observing System
EOSDIS	Earth Observing System Data and Information System
ESDIS Project	Earth Science Data and Information System Project
ESDS	Earth Science Data System (Program)
ESDSWG	Earth Science Data System Working Group
ESDT	Earth Science Data Type
ESO	ESDIS Standards Office
FAIR	Findable, Accessible, Interoperable, Reusable
GCMD	Global Change Master Directory
GeoTIFF	Georeferenced Tagged Image File Format

GES DISC	NASA's Goddard Earth Sciences Data and Information Services Center
GIS	Geographic Information System
GLDAS	Global Land Data Assimilation System
GPM	Global Precipitation Measurement (Mission)
GSFC	NASA Goddard Space Flight Center
GUI	Graphical User Interface
HDF5	Hierarchical Data Format, Version 5
HDF-EOS5	Hierarchical Data Format - Earth Observing System, Version 5 (based on HDF5)
HIRDLS	High Resolution Dynamics Limb Sounder
HPD	HDF Product Designer
ICARTT	International Consortium for Atmospheric Research on Transport and Transformation
ICD	Interface Control Document
IDL	Interactive Data Language
ISO	International Organization for Standardization
JPL	NASA Jet Propulsion Laboratory
KML	Keyhole Mark-up Language
L1, L2, L3	Level 1, Level 2, Level 3 (data product)
LAADS	Level-1 and Atmosphere Archive and Distribution System
LaRC	NASA Langley Research Center
MEaSUREs	**M**aking **Ea**rth **S**ystem Data Records for **U**se in **R**esearch **E**nvironment**s**
MERRA	Modern Era-Retrospective Analysis for Research and Applications
MODIS	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCO	NetCDF Operator
NetCDF-4	Network Common Data Form, Version 4
NOAA	National Oceanic and Atmospheric Administration
NRT	Near Real Time
NSIDC	National Snow and Ice Data Center
NUG	NetCDF Users Guide

OCO	Orbiting Carbon Observatory
OGC	Open Geospatial Consortium
ORNL DAAC	Oak Ridge National Laboratory DAAC (NASA)
PGE	Product Generation Executable
PO.DAAC	Physical Oceanography Distributed Active Archive Center (NASA JPL)
RFC	Request for Comments
SDP	Science Data Production (Toolkit)
SeaWiFS	Sea-viewing Wide Field-of-view Sensor
SEDAC	Socioeconomic Data and Applications Center (NASA)
SIPS	Science Investigator-led Processing System
THREDDS	Thematic Real-time Environmental Distributed Data Services
TRMM	Tropical Rainfall Measuring Mission
UDUNITS	A Unidata package that contains an extensive unit database
UMM	Unified Metadata Model
URL	Uniform Resource Locator
UTC	Coordinated Universal Time
uuid	Universal Unique Identifier
W3C	World Wide Web Consortium
WKT	Well-Known Text markup language
XML	eXtensible Markup Language

APPENDIX B. GLOSSARY

Calibration/Validation (Cal/Val) - Calibration is a demonstration that an instrument or device produces accurate results; validation is a program that provides assurance that a specific process, equipment, or system consistently produces results that meet pre-determined acceptance criteria.

Climate and Forecast Metadata Conventions - A set of metadata conventions that were invented for climate and weather forecast data but have since been applied to describe other kinds of Earth Science data, with the intention of promoting the processing and sharing of data.

Data Collection - A major release of a data product, or of a set of closely related data products, which can be followed by minor releases within the same collection.

Data Distributor - An entity responsible for archiving and distributing data products (e.g., a DAAC).

Data Processing Level – The level of processing that results in data products ranging from raw instrument data to refined analyses that use inputs from various sources [4].

Data Producer – A person or group that directly collects/creates data to be submitted to a NASA DAAC for archiving and public distribution.

Data Product - A set of data files that can contain multiple parameters, and that compose a logically meaningful group of related data.

Dataset - A broadly used term that can be used to describe any set of data. The official term “HDF5 dataset” describes a data array in an HDF5 file (equivalent to a NetCDF-4 variable in a NetCDF-4 file or an HDF-EOS field in an HDF-EOS file). At the opposite extreme, an entire data collection can also be referred to as a dataset.

Discovery – Successful identification and location of data products of interest.

Global Attribute – An attribute that applies to either the entire file or the entire collection of files. Some important examples are provided in Appendix D.

Granule - The smallest aggregation of independently managed (i.e., described, inventoried, retrievable) data at a DAAC. Some Web applications and services provided by DAACs allow for the subsetting of granules. One granule usually comprises one file, more rarely multiple files. The latter is not optimal as it complicates data management by both the archive and users, and utilization by tools and services.

Quality Flag - One or more unique variables within a data file that show what data quality assessments have been performed as well as diagnostics on various aspects of quality. A quality flag can be a byte value with each bit representing a pre-defined quality verification criterion provided as a Boolean expression. For example, see https://oceancolor.gsfc.nasa.gov/atbd/sst/flag/.

Quality Indicator - One or more unique variables within a data file whose numerical value shows the overall quality of a geophysical measurement. The numerical value should be on a pre-defined numerical scale. For example, uncertainty per pixel (or measurement), percent cloud cover (in a scene).

Search – An activity attempting to identify and locate data products of interest given user-defined criteria.

Self-Describing File – A file that contains metadata thoroughly describing the characteristics and content of the file.

UMM Profile – One of seven UMM metadata profiles: Collection (UMM-C), Granule (UMM-G), Service (UMM-S), Variable (UMM-Var), Visualization (UMM-Vis), Tools (UMM-T) and Common (UMM-Common).

Web Object Storage - Object storage accessible through https.

APPENDIX C. PRODUCT TESTING WITH DATA TOOLS

C.1 Panoply

Panoply [64], a tool developed and maintained by the NASA Goddard Institute for Space Studies, is particularly useful in verifying the proper geolocation of the data. If the geolocation has been set properly (i.e., horizontal extent), Panoply displays the georeferenced two-dimensional image (Geo2D) in the "Type" column of the Datasets Browser for maps and generates a map with boundary outlines and latitude and longitude grid lines (Figure 3).

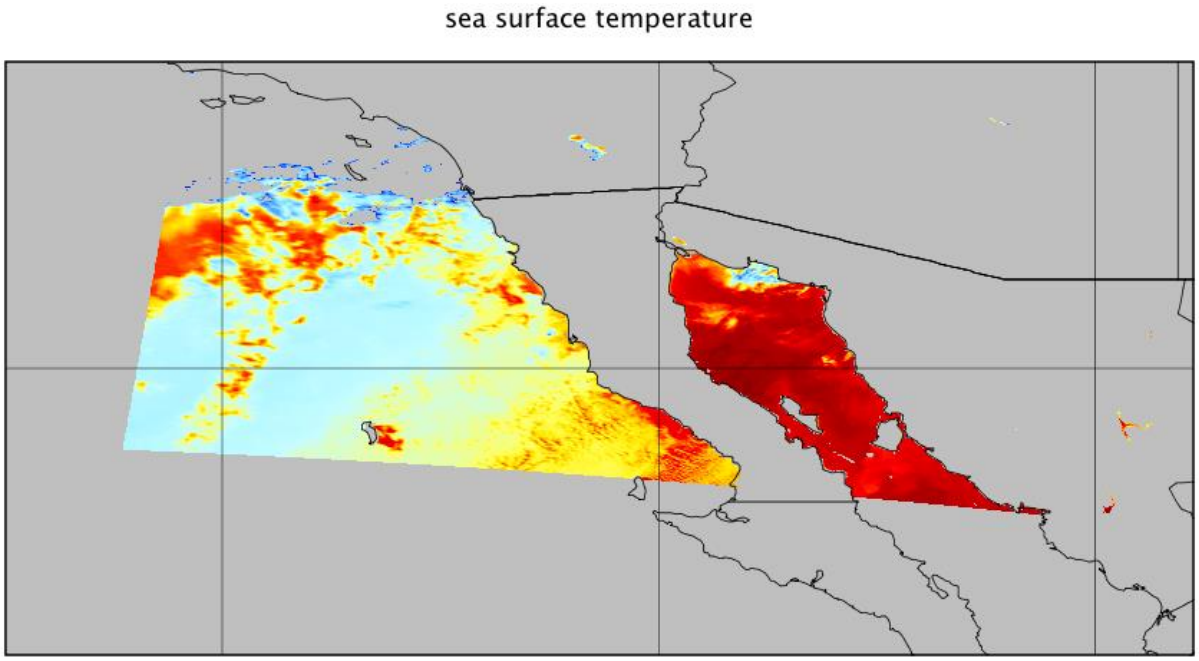

Figure 3. A screenshot of the Panoply software environment demonstrating a georeferenced two-dimensional image.

If the geolocation has not been set properly, it will display a two-dimensional image and generate a simple coordinate plot. Similarly, trajectory data will have type "GeoTraj" when properly geolocated; otherwise, it will display a one-dimensional image. Panoply also displays the variable-level and global metadata in the right-hand panel, making it convenient to confirm the map units and coordinate attributes.

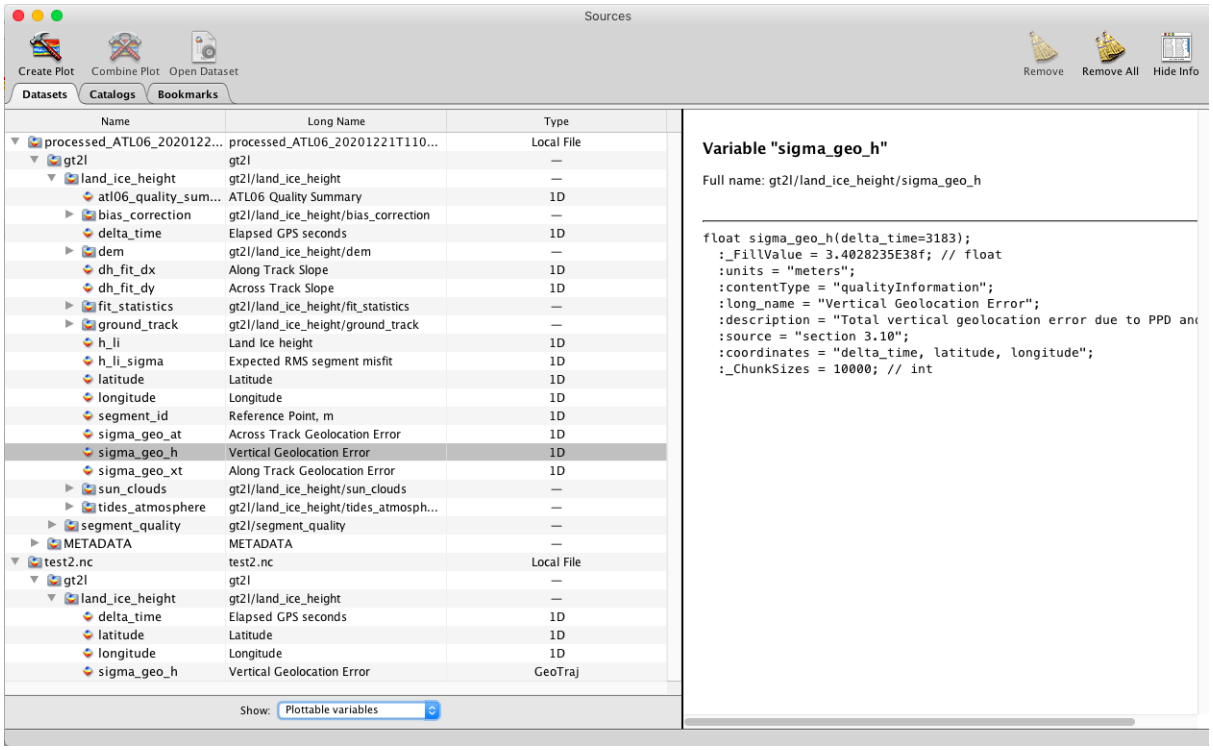

Figure 4. A screenshot of the Panoply software environment demonstrating a deviation from the CF Conventions.

Figure 4 illustrates that the variable “sigma_geo_h” for the data file “processed_ATL06_ … .nc” includes commas in the attribute `coordinates`, which violates the CF Conventions and causes Panoply to recognize the data as 1D. In the data file “test2.nc” there are no commas included in `coordinates`, Panoply recognizes the data as a GeoTraj, and the data can be plotted.

C.2 HDFView

HDFView from The HDF Group [63] is a tool for browsing and editing files in HDF and netCDF formats (e.g., Figure 5). Using HDFView, a user can view and modify the content of a file, view a file hierarchy in a tree structure, create new files, add or delete groups and data, and modify attributes. Unlike Panoply, the visualizations do not include a coastline overlay. However, the HDF-EOS plugin for HDFView can supply the latitude and longitude for each cell location in a data array.

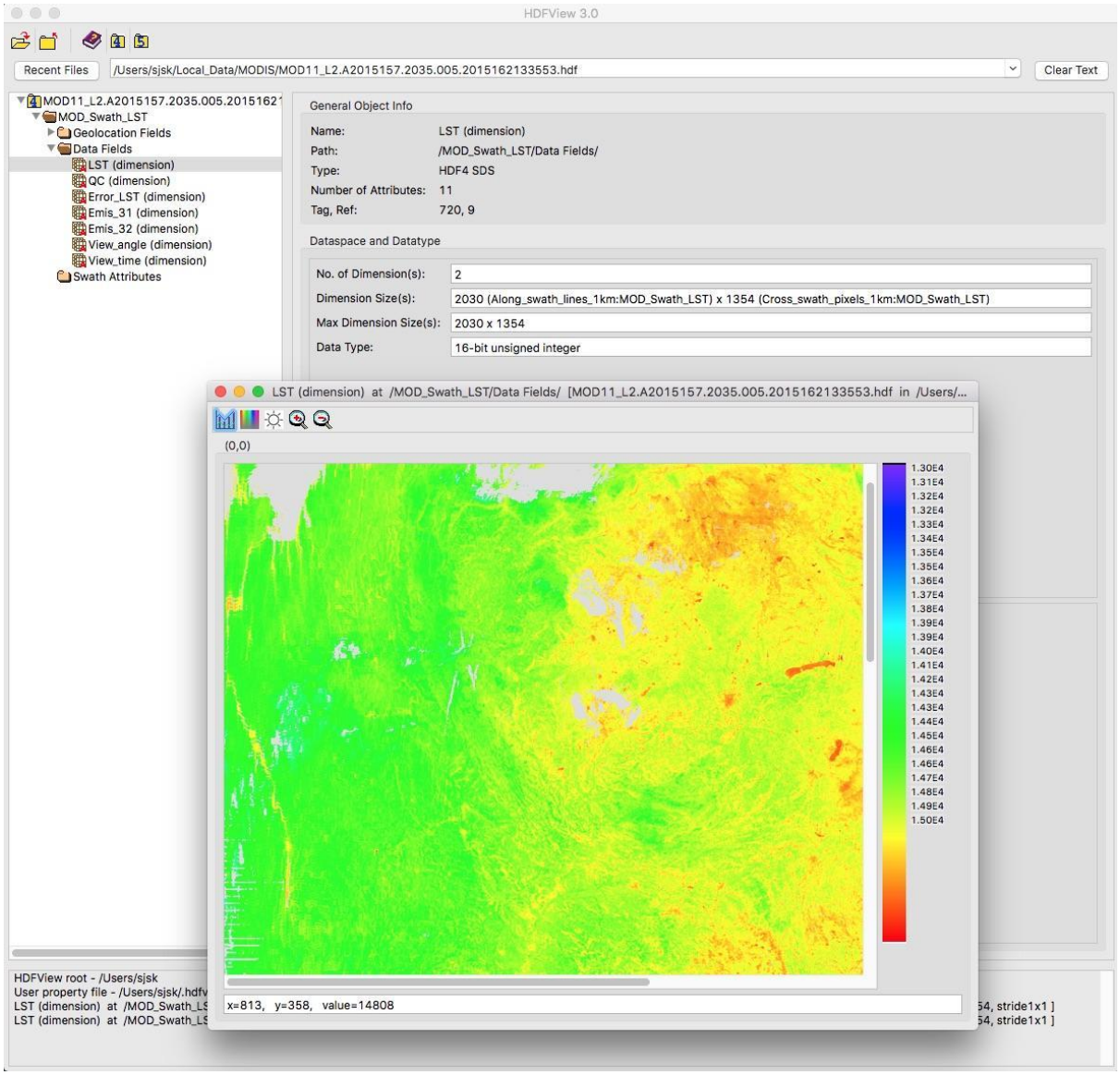

Figure 5. A screenshot of the HDFView software environment demonstrating how to view the dimension “LST” of a data swath.

APPENDIX D. IMPORTANT GLOBAL ATTRIBUTES

In this appendix, we provide a list of recommended attributes for the metadata headers at the global level according to CF (Version 1.7) [22] and ACDD (Version 1.3) [57], and other conventions. These are derived from the CF and ACDD Metadata Standards Overview by the Physical Oceanography Distribute Active Archive Center (PO.DAAC) [82]. The attributes where the indicated source is the Goddard Earth Sciences Data and Information Services Center (GES DISC) are from reference [5]. The attributes where the source is shown as Level-1 and Atmosphere Archive and Distribution System (LAADS) DAAC are from reference [83]. Where the PO.DAAC and GES DISC versions have conceptually similar attributes, the GES DISC names for the attributes are shown in parentheses in the Attribute Name column. It is to be noted that while there is considerable commonality in the attribute sets called for by the different DAACs, there are also some differences. Data producers are advised to consult the supporting DAAC for a list of required attributes (or others not shown in the list below). These recommended attributes have been characterized with a justification for use, and labelled with how they support *Findability, Accessibility, Interoperability, and Reusability (FAIR)* [84]. Also, the attributes whose names correspond to required elements indicated in [85] are marked with [R] in the tables below. The attributes are grouped based on the justification, i.e., what purpose they serve from the point of view of a data user and shown in Sections D.1 through D.6. Section D.6 shows the attributes needed for providing information on provenance and is divided into three subsections: General, Attribution, and Lineage. Where an attribute is considered to be important for more than one purpose, the justification column shows both the primary justification (corresponding to the section in which it appears) and the other purpose(s). It is to be noted that most of the listed attributes are at the data product level, while there are several that apply at the file level or at both levels. Data producers are advised to work with the DAACs to determine which attributes should be defined at which level, as well as where in the data products the attribute values should be stored.

D.1 Interpretability

Attribute Name	Definitions	Source	Justification	FAIR
title (LongName/Title) [R]	A short phrase or sentence describing the data. In many search systems, this attribute will be displayed in the results list from a search and should be human-readable and of a reasonable length.	CF 1.7, ACDD 1.3, NUG 1.7	Interpretability	F
summary (ProjectAbstract) [R]	A paragraph describing the data, analogous to an abstract for a manuscript.	ACDD 1.3	Interpretability	F
keywords_vocabulary	If using a controlled vocabulary for the words/phrases in keywords, this is the unique name or identifier of the vocabulary from which keywords are taken. If more than one keyword vocabulary is used, each can be presented with a prefix (e.g., “CF:NetCDF COARDS Climate and Forecast	ACDD 1.3	Interpretability	I

Attribute Name	Definitions	Source	Justification	FAIR
	Standard Names") and a following comma, so that keywords may optionally be prefixed with the controlled vocabulary key.			
Conventions (Conventions) [R]	A comma-separated list of the conventions that the data follow. For files that follow this version of ACDD, include the string "ACDD-1.3."	CF 1.7, ACDD 1.3, NUG 1.7	Interpretability	I
platform_vocabulary	Controlled vocabulary for the names used in platform.	ACDD 1.3	Interpretability	F, A
instrument_vocabulary	Controlled vocabulary for the names used in instrument.	ACDD 1.3	Interpretability	F, A
standard_name_vocabulary	The name and version of controlled vocabulary from which variable standard names are taken. Values for standard_name must come from the CF Standard Names vocabulary for the data to comply. Example: "CF Standard Name Table v27." Multiple distinct vocabularies can be separated by commas.	ACDD 1.3	Interpretability	I, R
Format	Format of the data, e.g., HDF-EOS5 or netCDF-4.	GES DISC	Interpretability	I
MapProjection	Applies to gridded data. Useful for application data tools, such as ArcGIS.	GES DISC	Interpretability	I
DataSetLanguage	Language used within the data set, LanguageCode=ISO 639 default=English.	GES DISC	Interpretability	I, R

D.2 Discovery

Attribute Name	Definitions	Source	Justification	FAIR
platform (Source) [R]	Name of the platform that supported the sensor data used to create the data. Platforms can be of any type, including satellite, ship, station, aircraft, or other. Indicate controlled vocabulary used in platform_vocabulary.	ACDD 1.3	Discovery	F, A
instrument (Source) [R]	Name of the contributing instrument or sensor used to create the data. Indicate controlled vocabulary used in instrument_vocabulary.	ACDD 1.3	Discovery	F, A
processing_level (ProcessingLevel) [R]	A textual description of the processing (or quality control) level of the data.	ACDD 1.3	Discovery	F, A, I
keywords (ProductParameters/ Keyword) [R]	A comma-separated list of keywords and/or phrases. Keywords can be common words or phrases, terms from a controlled vocabulary (e.g., Global Change Master Directory [45]), or Uniform Resource Identifiers [86] for terms from a controlled vocabulary. See also keywords_vocabulary.	ACDD 1.3	Discovery	F, A, I
ShortName [R]	An abbreviated name of the product, limited to 30 characters. This should contain the version and be somewhat identifiable. A legend for determining the ShortName meaning should be included in the documentation. The shortname can only contain "_" in the name, no other non-numeric characters and be under 30 characters. The shortname should be in all capital letters. Using mixed case is discouraged.	GES DISC	Discovery	F
GranuleID [R]	For most Level 1 to Level 4 products, this is the file name. It is recommended to include descriptive information, version, and date in the file name.	GES DISC	Discovery	F, A
OrbitNumber	(For swath data – Level 1 & 2) Sequential number assigned to satellite orbits (integer).	GES DISC	Discovery, Geolocation	F
EquatorCrossingLongitude	(For swath data – Level 1 & 2) Longitude at which the satellite crosses the equator (decimal degrees).	GES DISC	Discovery, Geolocation	F
EquatorCrossingDate	(For swath data – Level 1 & 2) Date on which the satellite crosses the equator (YYYY-MM-DD).	GES DISC	Discovery, Geolocation	F
EquatorCrossingTime	(For swath data – Level 1 & 2) Time at which the satellite crosses the equator (UTC hh:mm:ss).	GES DISC	Discovery, Geolocation	F

Attribute Name	Definitions	Source	Justification	FAIR
StartLatitude	(For swath data – Level 1 & 2) Nadir latitude at start of swath (+ or – 90 degrees).	GES DISC	Discovery, Geolocation	F
StartDirection	(For swath data – Level 1 & 2) Orbit direction at StartLatitude (A = ascending; D = descending).	GES DISC	Discovery, Geolocation	F
EndLatitude	(For swath data – Level 1 & 2) Nadir latitude at end of swath (+ or – 90 degrees).	GES DISC	Discovery, Geolocation	F
EndDirection	(For swath data – Level 1 & 2) Orbit direction at EndLatitude (A = ascending; D = descending).	GES DISC	Discovery, Geolocation	F
NumberofOrbits	(For swath data – Level 1 & 2) Number of orbits for the swath (needed if a file has more than 1 orbit).	GES DISC	Discovery, Geolocation	F
StartOrbit	(For swath data – Level 1 & 2) The start orbit number (needed if a file contains multiple orbits).	GES DISC	Discovery, Geolocation	F
StopOrbit	(For swath data – Level 1 & 2) The stop orbit number (needed if a file contains multiple orbits).	GES DISC	Discovery, Geolocation	F
metadata_link	A URL that gives the location of complete metadata. A persistent URL is recommended.	ACDD 1.3	Discovery, Provenance	F, A, I, R
product_version (VersionID) [R]	Version identifier of the data as assigned by the data creator. For example, a new algorithm or methodology could result in a new version of a product.	ACDD 1.3	Discovery, Provenance	F, A, I

D.3 Geolocation

Attribute Name	Definitions	Source	Justification	FAIR
geospatial_bounds (SpatialCoverage, ObservationArea) [R]	Describes the data's 2D or 3D geospatial extent in OGC's Well-Known Text (WKT) Geometry format (see the OGC Simple Feature Access specification [87]). The meaning and order of values for each point's coordinates depends on the CRS. The ACDD default is "2D geometry" in the "EPSG:4326" CRS. The default may be overridden with geospatial_bounds_crs and geospatial_bounds_vertical_crs. EPSG:4326 coordinate values are "latitude (decimal degrees_north)" and "longitude (decimal degrees_east)," in that order. Longitude values in the default case are limited to the (-180, 180) range. Example: "POLYGON (40.26 -111.29, 41.26 -111.29, 41.26 -110.29, 40.26 -110.29, 40.26 -111.29)."	ACDD 1.3	Geolocation	A, I
geospatial_bounds_crs (SpatialCoverage, ObservationArea) [R]	The CRS of the point coordinates in geospatial_bounds. This CRS may be 2D or 3D, but together with geospatial_bounds_vertical_crs (if supplied), must match the dimensionality, order, and meaning of point coordinate values in geospatial_bounds. If geospatial_bounds_vertical_crs is also present, then only specify a 2D CRS. EPSG CRSs are strongly recommended. If this attribute is not specified, the CRS is assumed to be "EPSG:4326." Examples: "EPSG:4979" (the 3D WGS84 CRS), "EPSG:4047."	ACDD 1.3	Geolocation	A, I
geospatial_bounds_vertical_crs (SpatialCoverage, ObservationArea) [R]	The vertical CRS for the Z axis of the point coordinates in geospatial_bounds. This attribute cannot be used if the CRS in geospatial_bounds_crs is 3D. To use this attribute, geospatial_bounds_crs must exist and specify a 2D CRS. EPSG CRSs are strongly recommended. There is no default for this attribute when not specified. Examples: "EPSG:5829" (instantaneous height above sea level), "EPSG:5831" (instantaneous depth below sea level), "EPSG:5703" (NAVD88 height).	ACDD 1.3	Geolocation	A, I
geospatial_lat_min (SouthernmostLatitude) [R]	Describes a simple lower latitude limit that may be part of a 2D or 3D bounding region. geospatial_lat_min specifies the southernmost latitude covered by the data.	ACDD 1.3	Geolocation	A, I
geospatial_lat_max (NorthernmostLatitude) [R]	Describes a simple upper latitude limit that may be part of a 2D or 3D bounding region. geospatial_lat_max specifies the northernmost latitude covered by the data.	ACDD 1.3	Geolocation	A, I

Attribute Name	Definitions	Source	Justification	FAIR
geospatial_lat_units (SpatialCoverage) [R]	Units for the latitude axis described in geospatial_lat_min and geospatial_lat_max. These are presumed to be “degrees_north,” but other options from the UDUNITS package can be specified instead.	ACDD 1.3	Geolocation	A, I
geospatial_lat_resolution (LatitudeResolution) [R]	Information about the targeted spacing of points in latitude. Describing resolution as a number value followed by the units is recommended. For L1 and L2 swath data, this is an approximation of the pixel resolution. Examples: “100 meters,” “0.1 degree.”	ACDD 1.3	Geolocation	A, I
geospatial_lon_min (WesternmostLongitude) [R]	Describes a simple longitude limit and can be part of a 2D or 3D bounding region. geospatial_lon_min specifies the westernmost longitude covered by the data. See also geospatial_lon_max.	ACDD 1.3	Geolocation	A, I
geospatial_lon_max (EasternmostLongitude) [R]	Describes a simple longitude limit; may be part of a 2D or 3D bounding region. geospatial_lon_max specifies the easternmost longitude covered by the data. Cases where geospatial_lon_min is greater than geospatial_lon_max indicate the bounding box extends from geospatial_lon_max, through the longitude range discontinuity meridian (either the antimeridian for -180:180 values, or Prime Meridian for 0:360 values), to geospatial_lon_min. For example, “geospatial_lon_min=170” and “geospatial_lon_max=-175” incorporates 15 degrees of longitude (ranges 170 to 180 and -180 to -175).	ACDD 1.3	Geolocation	A, I
geospatial_lon_units (SpatialCoverage) [R]	Units for the longitude axis described in geospatial_lon_min and geospatial_lon_max. These are presumed to be "degrees_east," but other options from the UDUNITS package can be specified instead.	ACDD 1.3	Geolocation	A, I
geospatial_lon_resolution (LongitudeResolution) [R]	Information about the targeted spacing of points in longitude. Describing resolution as a number value followed by units is recommended. For L1 and L2 swath data, this is an approximation of the pixel resolution. Examples: “100 meters,” “0.1 degree.”	ACDD 1.3	Geolocation	A, I
geospatial_vertical_min (SpatialCoverage) [R]	Describes the numerically smaller vertical limit and can be part of a 2D or 3D bounding region. See also geospatial_vertical_positive and geospatial_vertical_units.	ACDD 1.3	Geolocation	A, I
geospatial_vertical_max (SpatialCoverage) [R]	Describes the numerically larger vertical limit and can be part of a 2D or 3D bounding region. See also geospatial_vertical_positive and geospatial_vertical_units.	ACDD 1.3	Geolocation	A, I
geospatial_vertical_resolution	Information about the targeted vertical spacing of points. Example: “25	ACDD 1.3	Geolocation	A, I

Attribute Name	Definitions	Source	Justification	FAIR
(DataResolution) [R]	meters."			
geospatial_vertical_units (SpatialCoverage) [R]	Units for the vertical axis described in geospatial_vertical_min and geospatial_vertical_max. The default is "EPSG:4979" (height above the ellipsoid, in meters), but other vertical coordinate reference systems may be specified. Note that the common oceanographic practice of using pressure for a vertical coordinate, while not strictly a depth, can be specified using the unit bar. Examples: "EPSG:5829" (instantaneous height above sea level), "EPSG:5831" (instantaneous depth below sea level).	ACDD 1.3	Geolocation	A, I
geospatial_vertical_positive (SpatialCoverage) [R]	Values include either "up" or "down." If "up," vertical values are interpreted as "altitude," with negative values corresponding to below the reference datum (e.g., under water). If "down," vertical values are interpreted as "depth," positive values correspond to below the reference datum. Note that if geospatial_vertical_positive is "down" (depth orientation), geospatial_vertical_min specifies the data's vertical location furthest from the Earth's center, and geospatial_vertical_max specifies the location closest to the Earth's center.	ACDD 1.3	Geolocation	A, I

D.4 Temporal Location

Attribute Name	Definitions	Source	Justification	FAIR
time_coverage_start (RangeBeginningDate, RangeBeginningTime) [R]	Describes the time of the first data point in the data. Use the ISO 8601:2004 date format, preferably the extended format as recommended in ACDD Section 2.6 [57].	ACDD 1.3	Temporal Location	F, A, I
time_coverage_end (RangeEndingDate, RangeEndingTime) [R]	Describes the time of the last data point in the data. Use ISO 8601:2004 date format, preferably the extended format as recommended in in ACDD Section 2.6 [57].	ACDD 1.3	Temporal Location	F, A, I
time_coverage_duration (TemporalRange) [R]	Describes the duration of the data. Use ISO 8601:2004 duration format, preferably the extended format as recommended in ACDD Section 2.6 [57].	ACDD 1.3	Temporal Location	F, A, I
time_coverage_resolution (TemporalRange) [R]	Describes the targeted time period between each value in the data. Use ISO 8601:2004 duration format, preferably the extended format as recommended in ACDD Section 2.6 [57].	ACDD 1.3	Temporal Location	A, I

D.5 Usability

Attribute Name	Definitions	Source	Justification	FAIR
cdm_data_type	The data type, as derived from Unidata's Common Data Model Scientific Data (CDM) types and understood by Thematic Real-time Environmental Distributed Data Services (THREDDS). This is a THREDDS dataType, and is different from the CF featureType, which indicates a Discrete Sampling Geometry file.	ACDD 1.3	Usability	A, I
comment	Miscellaneous information about the data not captured elsewhere.	CF 1.7, ACDD 1.3,	Usability	F, A, I, R
acknowledgement	A place to acknowledge various types of support for the project that produced the data.	ACDD 1.3	Usability	R
license	Provide the Uniform Resource Locator (URL) to a standard or specific license, enter "Freely Distributed" or "None," or describe any restrictions to data access and distribution in free text.	ACDD 1.3	Usability	R
DataSetQuality	Overall assessment of quality of data, including relevant articles. Short summary is preferred.	GES DISC	Usability	R
DataProgress [R]	Status of dataset.	GES DISC	Usability	R
SpatialCompletenessDefinition	Definition of a measure of data quality: e.g., the ratio of grid elements containing valid values to total number of grid elements.	GES DISC	Usability	R
SpatialCompletenessRatio	The data quality information: value for the above-defined measure.	GES DISC	Usability	R
FOVResolution	(For swath data – Level 1 & 2) Field-of-view resolution of sensor used to acquire the swath data (if multiple sensor, list FOVResolution of each sensor).	GES DISC	Usability	R

D.6 Provenance

D.6.1 General

Attribute Name	Definitions	Source	Justification	FAIR
uuid	A uuid (Universal Unique Identifier) is a 128-bit number used to uniquely identify some object or entity on the Web. Depending on the specific mechanisms used, a uuid is either guaranteed to be different or at least extremely likely to be different from any other uuid generated until 3400 A.D. Applied to identify the data product the file belongs to.	NASA ESDIS	Provenance (General)	F, A
date_created [R]	The date on which the current version of the data was created. Modification of values implies a new version; hence, this would be assigned the date of the most recent values modification. Metadata changes are not considered when assigning this attribute. The ISO 8601:2004 extended date format is recommended, as described in ACDD Section 2.6 [57].	ACDD 1.3	Provenance (General)	R
date_modified [R]	The date on which the data was last modified. Note that this applies just to the data, not the metadata. The ISO 8601:2004 extended date format is recommended, as described in ACDD Section 2.6 [57].	ACDD 1.3	Provenance (General)	R
date_issued [R]	The date on which the data (including all modifications) was formally issued (i.e., made available to a wider audience). Note that these apply just to the data, not the metadata. The ISO 8601:2004 extended date format is recommended, as described in ACDD Section 2.6 [57].	ACDD 1.3	Provenance (General)	R
date_metadata_modified [R]	The date on which the metadata was last modified. The ISO 8601:2004 extended date format is recommended, as described in ACDD Section 2.6 [57].	ACDD 1.3	Provenance (General)	R
ValidationData	Description of or reference on how the data were validated.	GES DISC	Provenance (General)	R

D.6.2 Attribution

Attribute Name	Definitions	Source	Justification	FAIR
Id (IdentifierProductDOI) [R]	An identifier for the data product, provided by and unique to its naming authority. The combination of naming_authority and ID should be globally unique, but it can be globally unique per se. It can be Web addresses, DOIs, meaningful text strings, a local key, or any other unique string of characters. It should not include whitespace characters.	ACDD 1.3	Provenance (Attribution)	F, A
naming_authority (IdentifierProductDOIAuthority)	The organization that provides the initial identifier for the data. The naming authority should be uniquely specified by this attribute. It is recommended to use Reverse Domain Name System (e.g., [88]) for the naming authority. URIs are also acceptable (e.g., "edu.ucar.unidata").	ACDD 1.3	Provenance (Attribution)	I
creator_name (ContactPersonName)	The name of the person (or other creator type specified by creator_type) principally responsible for creating the data.	ACDD 1.3	Provenance (Attribution)	F, R
creator_email (ContactPersonEmail)	The email address of the person (or other creator type specified by creator_type) principally responsible for creating the data.	ACDD 1.3	Provenance (Attribution)	F, R
creator_url (RelatedURL)	The URL of the person (or other creator type specified by creator_type) principally responsible for creating the data.	ACDD 1.3	Provenance (Attribution)	F, R
creator_type (ContactPersonRole)	Specifies the type of creator with one of the following: person, group, institution, or position. If this attribute is not specified, the creator is assumed to be a person.	ACDD 1.3	Provenance (Attribution)	F, R
creator_institution (ContactPersonAddress)	The creator's institution. The value should be specified even if it matches the value of publisher_institution or if creator_type is an institution.	ACDD 1.3	Provenance (Attribution)	F, R
institution (Institution)	The name of the institution principally responsible for originating the data.	CF 1.7, ACDD 1.3	Provenance (Attribution)	F, R
project	The name of the project principally responsible for originating the data. Examples: "PATMOS-X," "Extended Continental Shelf Project."	ACDD 1.3	Provenance (Attribution)	F, R
program	The overarching program of which the data belongs. A program consists of a set (or portfolio) of related and possibly interdependent projects that meet an overarching objective. Examples: "GHRSST," "NOAA CDR," "NASA EOS," "JPSS," "GOES-R."	ACDD 1.3	Provenance (Attribution)	F, I, R
contributor_name	The name of any individuals, projects, or institutions that contributed to the creation of the data. Can be presented as free text, or in a structured	ACDD 1.3	Provenance (Attribution)	F, R

Attribute Name	Definitions	Source	Justification	FAIR
	format compatible with conversion to NcML (e.g., insensitive to changes in whitespace, including end-of-line characters).			
contributor_role	The role of any individuals, projects, or institutions that contributed to the creation of the data. Can be presented as free text, or in a structured format compatible with conversion to NcML (e.g., insensitive to changes in whitespace, including end-of-line characters). Multiple roles should be presented in the same order and number as the names in contributor_names.	ACDD 1.3	Provenance (Attribution)	F, R
publisher_name [R]	The name of the person (or other entity specified by publisher_type) responsible for publishing the data to users with its current metadata and format.	ACDD 1.3	Provenance (Attribution)	F, R
publisher_email	The email address of the person (or other entity specified by publisher_type) responsible for publishing the data to users with its current metadata and format.	ACDD 1.3	Provenance (Attribution)	F, R
publisher_url (RelatedURL)	The URL of the person (or other entity specified by publisher_type) responsible for publishing the data to users with its current metadata and format.	ACDD 1.3	Provenance (Attribution)	F, R
publisher_type	Specifies type of publisher with one of the following: person, group, institution, or position. If this attribute is not specified, the publisher is assumed to be a person.	ACDD 1.3	Provenance (Attribution)	F, R
publisher_institution (ProcessingCenter)	The publisher's institution. If publisher_type is an institution, this attribute should have the same value as publisher_name.	ACDD 1.3	Provenance (Attribution)	F, R

D.6.3 Lineage

Attribute Name	Definitions	Source	Justification	FAIR
InputDataProductVersion	Input data version.	GES DISC	Provenance (Lineage)	I, R
InputDataProducts	Input data to the product of interest.	GES DISC	Provenance (Lineage)	I, R
history (History, ProductGenerationAlgorithm,	Provides an audit trail for modifications to the original data. This attribute is also in the NUG: "This is a character array with a line for each invocation	CF 1.7, ACDD 1.3	Provenance (Lineage)	R

Attribute Name	Definitions	Source	Justification	FAIR
ProductGenerationAlgorithmVersion)	of a program that has modified the dataset. Well-behaved, generic netCDF applications should append a line containing: date, time of day, username, program name, and command arguments." To include a more complete description, append a reference to an ISO Lineage entity (see NOAA EDM ISO Lineage guidance [89]).			
source (Source, InputOriginalFile)	The method of production of the original data. If it was model generated, this attribute should provide the model and its version. If it is observational, this attribute should characterize it. Examples: "temperature from CTD #1234," "world model v.0.1."	CF 1.7, ACDD 1.3	Provenance (Lineage)	R
references (References)	Published or Web references that describe the data or methods used to produce the data. Recommend are URIs (such as a URL or DOI) for manuscripts or other references.	CF 1.7, ACDD 1.3	Provenance (Lineage)	R
ProductionDateTime	Date and time the file was produced.	GES DISC	Provenance (Lineage)	A, R
OriginalFileVersion		GES DISC	Provenance (Lineage)	R
PGEVersion	Version of the PGE that the product was generated from.	LAADS DAAC	Provenance (Lineage)	R
PGE_Name	Value like "PGExxx" where xxx is a number assigned to the PGE.	LAADS DAAC	Provenance (Lineage)	R
ProcessingEnvironment	Information on the machine where the code was executed.	LAADS DAAC	Provenance (Lineage)	R
PGE_StartTime	Value reflects the time at which the processing was started. Should be a value like YYYY-MM-DD HH:MM:SS.	LAADS DAAC	Provenance (Lineage)	R
PGE_EndTime	Value reflects the time at which the processing was completed. Should be a value like YYYY-MM-DD HH:MM:SS.	LAADS DAAC	Provenance (Lineage)	R

APPENDIX E. IMPORTANT VARIABLE ATTRIBUTES

In this appendix, we provide a list of recommended attributes for metadata headers at the variable level according to CF (Version 1.7) [22] and ACDD (Version 1.3) [57] conventions by PO.DAAC [82]. The attributes listed by GES DISC, from reference [5], are shown in parentheses in the Attribute Name column. It is to be noted that while there is considerable commonality in the attribute sets called for by the different DAACs, there are also some differences. The attributes whose names correspond to required elements indicated in [85] are marked with [R] in the tables below. Data producers are advised to consult the supporting DAAC for a list of required (or other) attributes. These recommended attributes have been characterized with a justification for use, and labelled with how they support *Findability, Accessibility, Interoperability, and Reusability (FAIR)* [84].

Attribute Name	Definitions	Source	Justification	FAIR
long_name (long_name) [R]	A descriptive (i.e., human-readable) name for the variable. Avoid including acronyms, abbreviations, and units.	CF 1.7, ACDD 1.3	Interpretability	F
standard_name (long_name)	Reserved for names that are part of the CF. If no CF standard name is appropriate for the variable, this attribute should be excluded. However, the CF is continually evolving: to get a standard name added to the CF, propose one by emailing to cf-metadata@cgd.ucar.edu.	CF 1.7	Interpretability	F, I
coverage_content_type	An ISO 19115-1 code to indicate the source of the data, MD_CoverageContentTypeCode [90]. Examples: “image,” “thematicClassification,” “physicalMeasurement,” “auxiliaryInformation,” “qualityInformation,” “referenceInformation,” “modelResult,” “coordinate.”	ACDD 1.3	Interpretability	F, A, I, R
units (units) [R]	This attribute is required for all variables that represent dimensional quantities (see [26], Rec. 3.1). We recommend adhering to the CF on the use of the units attribute with the following clarifications: a unitless (i.e., dimensionless, in the physical sense) variable is indicated by excluding this attribute, unless appropriate physical units do exist, then use of dimensionless units identifiers is common practice in the target user community. Values of units should be supported by the UDUNITS-2 library [53]. A variable used in any context other than data storage must not define units (see [26], Rec. 3.3).	CF 1.7	Data Services	I
_FillValue (_FillValue)	Include only if the variable has missing values. The value should be the same as that of the variable. Using NaN (Not a Number) to specify this attribute is discouraged (see [26], Rec. 3.7). A data producer can create a separate array to document the reasons for missing values by using flag_values and flag_meanings.	CF 1.7	Data Services	I

Attribute Name	Definitions	Source	Justification	FAIR
coordinates	This is necessary if the variable has coordinates that are not the same as the dimension names of the data. In this case, the coordinates attribute identifies the auxiliary variables that contain geospatial or temporal coordinates. For example, variables on a trajectory often use the data acquisition time as a single dimension but have auxiliary coordinates that record the latitude and longitude for each datum in the main variable. If a variable "foo" has a coordinate of time but co-exists with variables named "lat" and "lon" with the corresponding geolocation information, then the coordinates variable would read "lat lon" to indicate that the variables hold the appropriate geospatial location information. Without this attribute, tools will not be able to locate the corresponding latitude and longitude for data in a swath, trajectory, or non-rectilinear grid. Note that the auxiliary coordinate variables must still follow the conventions for units that are noted in Section 4.4 and Section 4.5 above. Also, note that the only delimiter between words inside the coordinates attribute should be a space.	CF 1.7	Data Services	A, I, R
bounds	An attribute for coordinate variables to describe the vertices of the cell boundaries and thus the intervals between cells	CF 1.7	Data Services	A,I,R
scale_factor (scale_factor)	This is often used to represent floating point numbers as short integers, thus resulting in more compact data (i.e., packed data). To convert the short integer value, it is multiplied by scale_factor and then add_offset is added. These attributes should be floating point numbers, not strings, to work properly. If the scale_factor is "1.0" and add_offset is "0.0," these attributes are omitted.	CF 1.7	Data Services	I
add_offset (scale_offset)	This is often used to represent floating point numbers as short integers, thus resulting in more compact data (i.e., packed data). To convert the short integer value, it is multiplied by scale_factor and then add_offset is added. These attributes should be floating point numbers, not strings, to work properly. If the scale_factor is "1.0" and add_offset is "0.0," these attributes should be omitted.	CF 1.7	Data Services	I
valid_min	A scalar specifying the minimum valid value for the variable.	CF 1.7	Data Services	I
valid_max	A scalar specifying the maximum valid value for the variable.	CF 1.7	Data Services	I
valid_range (valid_range)	A vector of two numbers specifying the minimum and maximum valid values for the variable, equivalent to specifying values for both valid_min and valid_max attributes. The attribute valid_range should not be defined if either valid_min or valid_max are defined.	CF 1.7	Data Services	I

Attribute Name	Definitions	Source	Justification	FAIR
grid_mapping	Describes the horizontal coordinate system. This attribute should indicate a variable that contains the parameters corresponding to the coordinate system. There are typically several parameters associated with each coordinate system. The CF defines a separate attribute for each of the parameters. Examples: “semi_major_axis,” “inverse_flattening,” “false_easting.”	CF 1.7	Data Services	I, R
flag_values (flag_values)	An enumerated list of status flags indicating unique conditions whose meaning is described by the commensurate list of descriptive phrases in attribute flag_meanings. The status flags are scaler of the same type as the described variable.	CF 1.7	Quality Filtering	I
flag_masks	A number of independent Boolean (i.e., binary) conditions using bit field notation and setting unique bits whose values are associated with a list of descriptive phrases in attribute flag_meanings. This attribute is the same type as the variable and contains a list of values matching unique bit fields.	CF 1.7	Quality Filtering	I
flag_meanings (flag_meanings)	A list of strings that define the physical meaning of each flag_masks bit field or flag_values scaler field. The strings are often phrasing with words concatenated with underscores, and strings are separated by a single space. The CF allows a single variable to contain both flag_values and flag_masks. In such cases, the interpretation of the flags is slightly tricky. flag_masks is used to “group” a set of flag_values into a nested conditional. See [22], Section 3.5 on how to interpret flag_meanings. It is recommended that Boolean (i.e., flag_masks) and enumerated flags (i.e., flag_values) be kept in separate variables.	CF 1.7	Quality Filtering	I
comment (comments)	Provides the data producer an opportunity to further describe the variable and inform the user of its contents via a text statement.	CF 1.7	Usability	I, R